# 通过关键示例定位的多示例学习算法用于乳腺微钙化簇检测

李超[1,2]，Kin Man Lam[2]，张磊[3]，惠春[1]，张素[1]

[1]生物医学工程学院，上海交通大学，上海
[2]信号处理中心，通讯与资讯工程系，香港理工大学，香港
[3]计算机系，香港理工大学，香港

## 摘要

本文提出一个基于多示例学习算法（Multi-Instance Learning，MIL）的计算机辅助诊断方案用于检测乳腺微钙化簇（microcalcification clusters，MCCs）。为了得到令人满意的检测结果，我们首先通过均值漂移算法来寻找乳腺 X 射线图像上可能的微钙化点疑似点，然后基于图构建的方式来提取单个疑似点的特征，最后本算法通过定位关键示例算法的多示例学习算法对这些疑似点进行分类。本实验结果在公共 DDSM 数据库上取得了很好的结果。

## 关键词

特征，微钙化簇，多示例学习，图，均值漂移

## 1 引言

很多实际应用中获得的数据都有其内在结构，相比于传统的单示例学习，多示例学习显著的优势之一在于更加自然地对目标进行表示且获得更多信息。例如，如果将一幅图像分割成多个部分，每一部分相对于将整幅图像作为一个示例，整幅图像作为一个包，那么多示例表示方法就可以捕获各部分的信息。若这个分割是有意义的（分割的每一部分对应某个主要特点），则多示例表示就有助于简化学习任务。

多数关于多示例学习算法都假设包中的示例是独立同分布的（independently and identically distributed，I.I.D），这样的假设会忽略示例之间的关系所显示的重要的结构信息。Zhou 和 Xu 等人指出应当假设包之间是服从独立同分布的，而并不假设包中的示例也服从独立同分布更具合理性。在实际应用中，包中的示例很少是独立存在的，因此在本节中我们假设包中示例是非独立同分布的（non-i.i.d）。Zhou 和 Sun 在后续研究工作中提出一种解决方法，该方法的基本思想是将包看做一个整体，包中的示例看做这个整体的组成部分，故引入图的概念。不同于 McGovern 和 Jensen 的研究工作中所处理的关系数据中每个示例本身就是一幅图，很多给定的数据不存在天然的图，因此需要根据数据构建图。

图（graph）是模式识别中用于对象表示的一种常规数据结构。对象的单个部分可以用图的节点（node）表示，图的边（edge）用于表示节点间的二元关系。图的节点和边都可以标有属性值，如表示成特征向量的形式，也是最常见的形式。图以其强大的表示能力在模式

识别领域取得很大的成功，如生物信息学（bioinformatics）、图像分类（image classification）和计算机网络分析（computer network analysis）等。

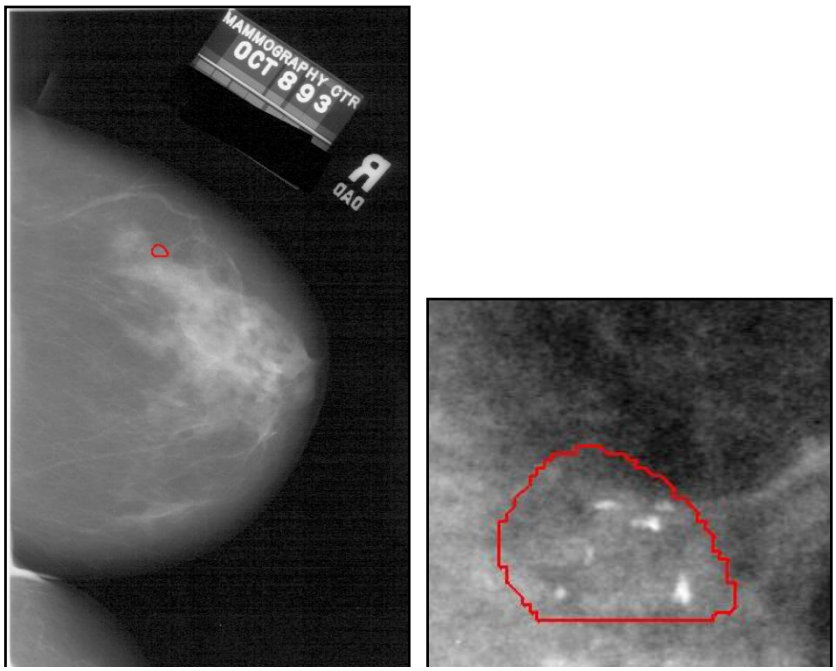
根据美国癌症协会（American Cancer Society）2013~2014 年的乳腺癌事实与数据年度报告，2013 年约有 23 万新增女性病例被确诊为浸润性乳腺癌，以及另外约有 65 万病例确诊为原位乳腺癌，并且呈现逐年上升的趋势。多数乳腺癌都具有浸润性，这些癌变组织可以从原位穿透导管或者腺体壁进而感染周围的乳腺组织。

浸润性乳腺癌的预后（prognosis）主要受到该疾病所处阶段的影响，即第一次确诊时的癌症程度或者扩散程度。当乳腺肿瘤比较小时无任何症状，同时也容易治愈。因此早期诊断对于乳腺癌患者来说是至关重要的，这也说明了乳腺筛查的必要性。直到现在，乳腺 X 射线摄片（mammography）筛查都是乳腺癌早期检测最方便可靠的检测方式。乳腺 X 射线摄片筛查过程中图像的阅读诊断是一项繁重的工作，一方面由于筛查的乳腺图片通常会受到图像质量和放射科医生专业水平的影响；另一方面筛查得到乳腺图像数量之多，也给放射科医生造成巨大的工作负担。鉴于这些因素，放射科医生检测可疑异常区域容易出现错误，有研究表明在筛选阶段会有 10%到 30%的放射科医生会做出错误的判断。为了减少此种错误和降低医生的工作量，自动进行乳腺癌初期筛查或者计算机辅助乳腺癌诊断就显得非常重要。

乳腺癌诊断的重要症状是乳腺微钙化簇（microcalcification clusters，MCCs），肿块和结构的扭曲紊乱。乳腺钙化是乳腺软组织内的钙沉积物，分为两种类型：大钙化灶和微钙化灶。通常情况下，微钙化是癌症的前兆，乳腺 X 射线图像上的微钙化簇被认为是乳腺癌最重要的标志之一。然而，微钙化点尺寸小、对比度低，同乳腺 X 射线图像上某些致密组织相似，这些特点增加了检测的难度。目前已有商用的计算机辅助诊断系统出现，并被放射科医生广泛使用。然而现有的 CAD 系统在筛查过程中对假阳性样本的判断仍然不尽如人意，因此许多研究指出 CAD 用于微钙化簇检测系统的设计依然是一个开放的研究领域。

为了对大量的乳腺 X 射线图像进行阅读诊断，用于自动检测微钙化簇的高效准确算法的设计是不可缺少的。Bozek 等人在乳腺数字 X 射线图像的图像处理算法进行了综述。Carlos 等人对乳腺 X 射线图像 CAD 系统中主要分类算法，即 SVM 和神经网络算法在乳腺公共临床数据库上进行了对比实验。De Santo 等人提出多分类器系统，系统设计的第一个分类器用于单个微钙化点的分类，第二个分类器用于簇的分类。其中单个钙化点所提取的特征包括紧实性、粗糙度、边界梯度强度和局部对比度等。Oliver 等人通过一组滤波器提取微钙化点形态学的局部描述获得局部特征，而后通过初始训练阶段学习选择重要特征，然后通过 boosted 分类器检测单个微钙化点进而检测微钙化簇。

Krishnapuram 等人最早将多示例学习算法引入到乳腺肿块的计算机辅助检测中，作者在文中指出，现成的 CAD 的核心算法，如 SVM、NN 算法等都有潜在的假设，即训练集合中的数据是独立同分布的。但是对于某些乳腺数据而言，同一乳腺同一图像的不同区域或者同一乳腺不同图像之间存在相关性，这种情况下训练数据就是非独立的。进一步来说，常规的检测算法最大化单个疑似区域的分类准确率，在实际情况中，按照每个病人或者每幅图像检出的准确率更加合理。作者指出基于多示例学习所提出的算法能够解决上述问题，同时还能够降低时间复杂度，算法更加有效准确。




a b

图 1 a. DDSM 数据库一张原始图片示例，并且图片有放射科医生勾画标注的金标准（ground truth）；b.勾画区域的放大

Fig.1 a. An example of a raw DDSM mammogram with a radiologist labeled ground truth (red curve). b. Enlargement of the ground-truth region.

根据我们文献调研的结果，目前尚无文献利用多示例学习算法进行乳腺微钙化簇的检测。现有的用于 MCCs 检测的机器学习算法通常需要对每一个疑似微钙化点的区域（candidate）进行标记，但是这些标记工作即使对于非常有经验的医生也是很大的挑战，如图 1 所示，图 a 为本章实验中 DDSM 数据库中的乳腺图像，b 图为微钙化簇区域放大的图像，红色曲线是有经验的放射科医生标记的含有微钙化簇区域的金标准（ground truth）。放射科医生很难给出对应于每个点标记结果，这就给那些对单个疑似点进行分类的有监督分类器的应用带来很大不便，标签的不准确会导致分类结果的不准确。因此在本章中，引入多示例学习算法，在没有单个疑似点（示例）标签，仅有候选区域（包）标签的情况下，进行特征提取并分类。

## 2 基于图特征的 KI-SVM 多示例学习方法

本实验计算机辅助诊断的流程框图如图 2 所示。首先滑动窗在待诊断图像上滑动，然后针对滑动窗内的每个候选点提取一组特征，一个滑动窗内特征向量组成一个包。多示例学习分类器将包分类为正常组织或者乳腺微钙化簇（MCC）。具体的计算机诊断方案的步骤会在后文详细介绍。

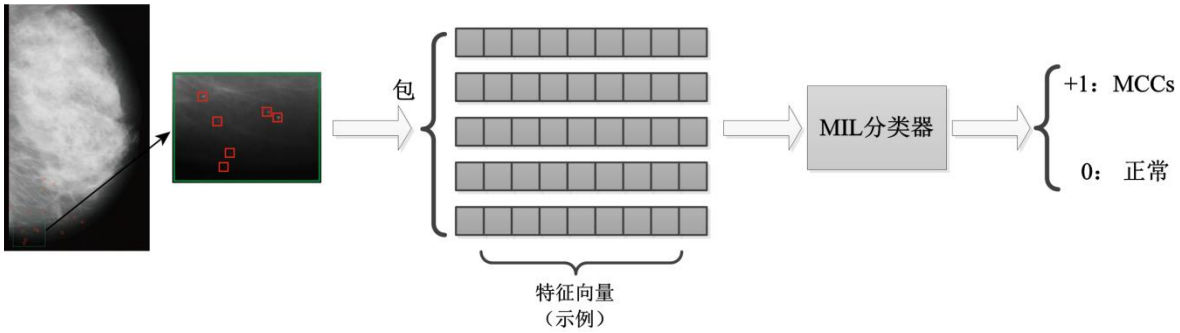



图 2 计算机辅助诊断乳腺 X 射线图片微钙化簇实验流程框图。首先滑动窗在待诊断图像上滑动，然后滑动窗内的每个候选点提取一组特征，一个滑动窗内特征向量组成一个包。多示例学习分类器将包分类为正常组织或者乳腺微钙化簇（MCC）。

Fig.2 The block diagram of the used computer-aided diagnosis pipeline. The silding window is first sliding on diagnostic image to extract bags, and then a set of features are extracted from each candidate inside a bag, and an instance is formed by the resultant feature vector. The bag is then classified into MCC or Normal by a multiple instance classifier.

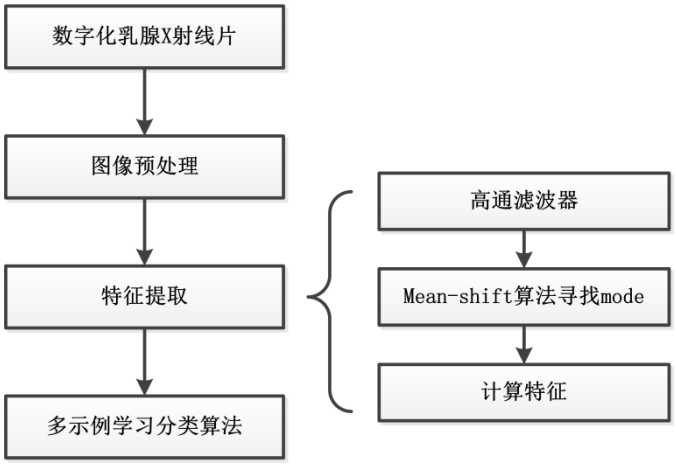



图 3 数字化乳腺 X 射线图像计算机辅助检测微钙化簇算法流程图

Fig.3 Flow chart of our proposed algorithm for computer aided detection of MCCs in digitized mammograms.

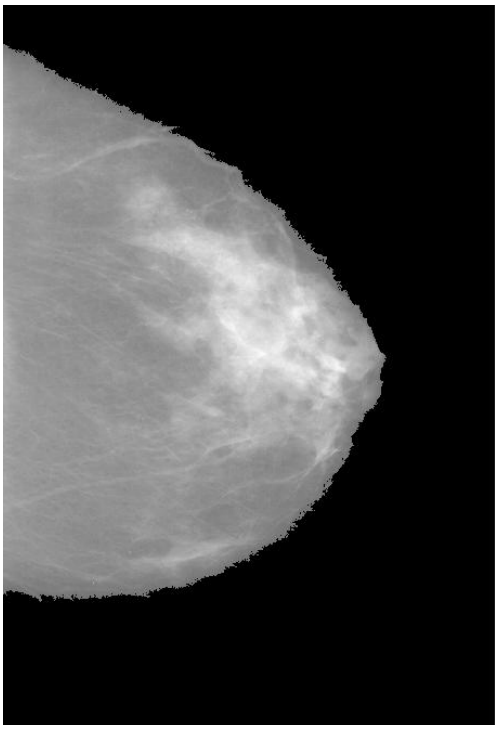


**图 4 原乳腺图像经过去噪和降采样后所提取的乳腺区域**

Fig.4 An enhanced and extracted breast region after denoising and downsampling

本文中，我们将提出的算法在 digital database of screening mammography (DDSM)公共数据库上进行测试。DDSM 数据库为研究学者提供无损压缩的数字化乳腺 X 射线图像的原始像素值，放射研究学者提供的病灶描述、评估和活检病理结果。本章提出用于乳腺 X 射线图像微钙化簇检测的基于多示例学习算法流程图（图 3）。

从 DDSM 数据库获取的乳腺图像首先经过高斯滤波器滤波除去冲击噪声。各向同性的高斯（Gaussian）函数如式（1）所示：

$$G(x,y) = \frac{1}{2ps^2} e^{-\frac{x^2+y^2}{2s^2}} \qquad (1)$$

数据库中原始图像大小约为$4000 \times 4000$，造成很高的计算复杂度，因此我们通过降采样方法将图像减小到约原图大小的四分之一。然后利用设定的阈值和形态学操作算子提取乳腺部分（如图 4 所示）。预处理的最后一步是经高通滤波器滤波后获得图像的高频成分，这是由于钙化点边缘同乳腺实质不同，属于图像中高频成分。

## 2.1 基于 mean-shift 算法生成候选点

图像经过预处理后，我们采用 256×256 像素大小的滑动窗（sliding window）在图像上滑动来选取 ROI，相邻两个滑动窗重叠 64 个像素。接下来每个滑动窗选取区域内通过 mean-shift 算法寻找势点（mode）。在计算机视觉领域中，Mean-shift 是一种寻找势点的算法，属于非参数方法，能够从复杂的多峰特征空间根据数据分布的概率密度函数估计鞍点（势点）。

Mean-shift 算法是通过迭代在给定分布中寻找局部灰度极值。假设给定数据集 $x_i \in \mathbb{R}^d$ 中 $n$ 个点，多变量核密度值 $\hat{f}_K$ 可以通过径向对称核函数 $K(x)$（如高斯核函数）求得，即式（2）和式（3）：

$$\hat{f}_K = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2}$$

$$K(x) = c_k k\left(\|x\|^2\right) \tag{3}$$

其中 $h$ 为核的半径值。

令 $g(x) = -k'(x)$ 表示所选核的导数，然后密度估计量的梯度可以通过式（4）计算：

$$m(x) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left\|\frac{x - x_i}{h}\right\|^2\right]}_{\text{第一项}} \underbrace{\left[\frac{\sum_{i=1}^{n} x_i g\left\|\frac{x - x_i}{h}\right\|^2}{\sum_{i=1}^{n} g\left\|\frac{x - x_i}{h}\right\|^2} - x\right]}_{\text{第二项}} \tag{4}$$

上式的第一项同 $x$ 处的密度估计成比例，第二项称为均值漂移向量，记为 $m(x)$，即：

$$m(x) = \frac{\sum_{i=1}^{n} x_i g\left\|\frac{x - x_i}{h}\right\|^2}{\sum_{i=1}^{n} g\left\|\frac{x - x_i}{h}\right\|^2} - x \tag{5}$$

均值漂移向量总指向密度值增加的方向。势点检测可以总结为以下步骤：

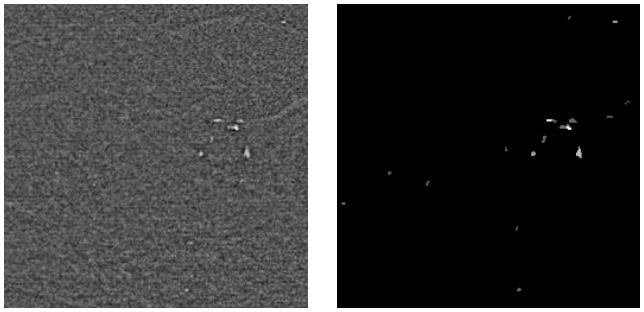
(1) 在整个特征空间内利用 mean-shift 算法寻找所有的鞍点（势点）；

(2) 保留局部最大值，删除其它鞍点，然后将距离小于某阈值内的势点进行合并；

(3) 对剩余的势点进行聚类，即可找到以每个势点为吸引中心的任意形状的聚类区域。

## 2.2 基于图的特征提取

通过上段所述 mean-shift 算法寻找得到的势点，我们在每个包（滑动窗覆盖的区域）中选取前 $t$（本实验取值 20）个最大的势点值对应的像素作为示例的中心点。如图 5 所示，图 a 为预处理后得到的图像上滑动窗覆盖的某区域（包），图 b 为 mean-shift 算法选取前 $t$ 个最大势点对应的疑似微钙化点（示例）。在这些得到的示例中，有些是钙化点，有些是较小且致密的乳腺区域，也有一些是冲击噪声。其中真正钙化点就是我们认为的正示例，其余的则是标签为负的示例。



a b

图 5 a.高通滤波结果, b.滑动窗内产生的候选点（包中的示例）

Fig.5 a. Result after high-pass filtering a candidate region, and b. the generated candidates within a sliding window (the instances within a bag).

表 1 所提取的示例特征

Table 1 The features used to describe a region

| index | 示例特征 |
|---|---|
| **1** | Area(number of pixels) |
| **2** | Mean Intensity |
| **3** | Eccentricity |
| **4** | Major Axis length |
| **5** | Degree |

绝大多数用于检测 MCCs 的机器学习算法通常用大量的特征做训练，这样会大大增加计算量和计算时间。本章提出的算法只用了 5 个特征，可以大大减少特征间的冗余和计算量。表 1 中列出了提取的特征，其中前两个特征分别表示每个示例的像素值和平均灰度值。第 3 个和第 4 个特征是将每一示例看作一个椭圆，分别计算其二阶偏心率和长轴的长度。第 5 个特征用于提取示例间的空间信息。

我们认为示例之间有比较强的空间关系，基于此通过创建图来描述示例之间的关系。假设对于包 $X_i$，其中的两个示例 $x_{ii}$ 和 $x_{ij}$，$x_{ij}, x_{ij} \hat{I} X_i$，计算他们之间的欧式距离，如果 $x_{ii}$ 和 $x_{ij}$ 之间的距离小于某个预先设定的阈值 $s$，则在两个节点间定义一条边，这样每个节点的度（边的条数）记为第 5 维特征——度（degree）。

## 2.3 基于关键示例定位的多示例学习算法

在乳腺 X 射线图像中检测 MCCs 预处理以及特征提取过程中，根据对其特点的分析，微钙化簇中存在一个或者几个关键微钙化点对检测起着重要的作用，也就是说这一个或者这几个微钙化点对包的标签确定起着重要的作用，因此我们认为定位关键的示例对于辨别包的标签是非常重要的。借鉴 Zhou 等人提出的基于关键示例定位的 KI-SVM 算法，具体算法如下：

令 N 为包的总数目，$p$ 为正包的数目。包的标签 $y_i = \begin{cases} 1 & 1 \le i \le p \\ 0 & p < i \le N \end{cases}$。引入指示变量 $d$ 表示包内示例是否为关键示例（key instance），并假设对于某标签为正的包 $X_i$，二值向量 $d_i = (d_{i,1}, \cdots, d_{i,m_i})^T \in \{0,1\}^{m_i}$ 表示 $X_i$ 中的关键示例。鉴于对乳腺微钙化点特点的分析和所提取的特征，可以认为只要其中有一个正的示例（定义为处于某微钙化簇中的钙化点），则该包的标签为正，记为 $\sum_{j=1}^{m_i} d_{i,j} = 1$

示例水平的 Ins-KI-SVM 可以写成式（6）：

$$\min_{w,r,x,d} \quad \frac{1}{2}\|w\|_2^2 - r + \frac{C}{2}\sum_{i=1}^{p} x_i^2 + \frac{\lambda C}{2}\sum_{i=p+1}^{m}\sum_{j=1}^{m_i} x_{s(i,j)}^2 \tag{6}$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} w\phi(d_{i,j} f(x_{i,j})) \ge r - x_i, \; i = 1, \cdots, p$$

$$-w\phi(f(x_{i,j})) \ge r - x_{s(i,j)}, \; i = p+1, \cdots, r$$

$$j = 1, \cdots, m_i$$

其中 $x_{i,j}$ 是第 $i$ 个包的第个 $j$ 示例。$C$ 是正则化参数，$x$ 为松弛变量，$\lambda$ 为用于平衡正包和负包的松弛变量。$j$ 为特定核函数的特征映射。

包水平的定位关键示例的算法 Bag-KI-SVM 的目标函数可以写成：

$$\min_{w,r,x,d} \quad \frac{1}{2}\|w\|_2^2 - r + \frac{C}{2}\sum_{i=1}^{p} x_i^2 + \frac{\lambda C}{2}\sum_{i=p+1}^{r} x_i^2 \tag{7}$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} w\phi(d_{i,j} f(x_{i,j})) \ge r - x_i, \quad i = 1, \cdots, p$$

$$-w\phi\left(\sum_{j=1}^{m_i} f(x_{i,j})\right) \ge r - x_i, \quad i = p+1, \cdots, r$$

上式中第二个约束项是对负包的约束，没有对负包内单个示例的约束。$C$ 是正则化参数，$x$ 为松弛变量，$\lambda$ 为用于平衡正包和负包的松弛变量。$j$ 为特定核函数的特征映射。

## 3 实验

实验数据为 DDSM 数据库的一个子集，数据库中图像分辨率比较高（50 或 43.5 微米每像素），像素水平的微钙化簇已经被放射学专家手动标记出来作为金标准。从 DDSM 数据库中选择了 40 幅含有 MCCs 的乳腺图像，经过预处理生成了 40 个正包、990 个负包，共 1130 个包。由于在生成候选点时经常会产生数以千计的候选点，计算所有候选点的特征就非常耗时，因此分类过程如果在保证敏感性不下降的情况下使用尽量少的特征是非常必要的，

本章所提出算法针对每个包提取 5 维特征组成紧致特征集。

**表 2 不同算法的曲线下面积**

Table 2 The AUCs of different algorithms.

| 算法 | 曲线下面积（AUC） |
|---|---|
| **Iterated APR** | 0.5313 |
| **DD** | 0.7580 |
| **EM-DD** | 0.7435 |
| **Citation-KNN** | 0.5861 |
| **MI-SVM** | 0.6448 |
| **mi-SVM** | 0.5599 |
| **Inst-KI-SVM** | **0.8775** |
| **Bag-KI-SVM** | **0.8618** |

在这些包中，进行 20 次重复试验得到的结果取均值，每次试验随机选取 20 个正包和 80 个负包作为训练集，剩下的包作为测试集。本章中检测结果测量指标有敏感性（sensitivity），特异性（specificity）和曲线下面积（Area Under Curve，AUC）。敏感性和特异性分别通过下式计算：

$$sensitivity = \frac{TP}{TP+FN}$$

$$specificity = \frac{TN}{TN+FP}$$

其中 TP 表示真阳性样本，FP 表示假阳性样本，TN 表示真阴性样本，FN 表示假阴性样本。AUC 表示接受者操作特征（receiver operating characteristic，ROC）曲线下面积值。

图 6 描述基于定位关键示例的 bag-KI-SVM 和 inst-KI-SVM 两种算法的 ROC 曲线。结果显示两种方法在 DDSM 数据库上取得了更高的准确率。多示例学习库 MILL (Multiple-Instance Learning Library)是一个开源的多示例学习算法工具包[12]。我们将 MILL 库中的经典基准算法同 KI-SVM 算法进行比较，实验结果见表 2 以及图 7。多示例学习算法：Iterated APR、DD、EM-DD、Citation-KNN、MI-SVM、mi-SVM、Inst-KI-SVM 和 Bag-KI-SVM 的曲线下面积分别为 0.5313、0.7580、0.7435、0.5861、0.6448、0.5599、0.8775 和 0.8618。实验结果表明通过定位关键示例的多示例学习算法能够获得更好的结果。证明了提取的基于图的特征的有效性，能够更好地描述微钙化簇中钙化点的性质。

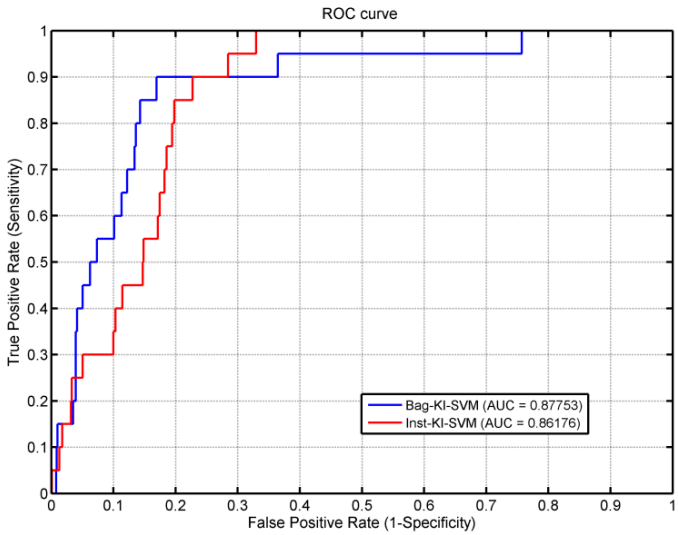



图 6 多示例学习算法 Bag-KI-SVM 和 Inst-KI-SVM 的 ROC 曲线

Fig.6 The ROC curves of Bag-KI-SVM and Inst-KI-SVM with $\mathrm{AUC}_{bag} = 0.87753$ and $\mathrm{AUC}_{inst} = 0.86176$, respectively.

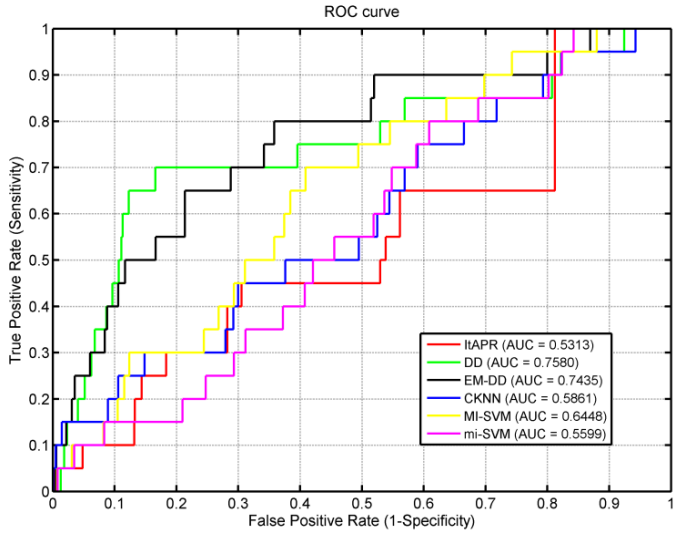



图 7 多示例学习算法 Iterated APR、DD,EM-DD、Citation-KNN、MI-SVM、mi-SVM 的 ROC 曲线

Fig.7 The ROC curves of the Iterated APR, DD, EM-DD, Citation-KNN, and MI-SVM algorithms (with the AUC listed in Table 2).

## 4 结论

本文针对乳腺 X 射线图像中的计算机辅助微钙化簇检测的问题，引入图论中思想，基于图表示从特征提取和分类算法两个层面上分别提出解决方案。首先提出一种新的基于图的

特征提取结合定位关键示例的多示例学习算法的计算机辅助检测方案。不同于以往多种计算机诊断方案需要提取大量特征，本章提取 5 维基本特征，其中基于图的特征提取很好的揭示了包中示例之间的内在联系，所得到的紧致特征集结合基于定位关键示例的多示例学习算法，在公共 DDSM 乳腺数据库上取得了较好的检测结果。

## 致谢

本文工作受到香港理工大学内部项目（Project No. G-U975）和中国国家基础研究计划 (973 Program, No. 2010CB732506)。

## 参考文献

[1] American Cancer Society. “Breast Cancer Facts & Figures 2011-2012”. Atlanta: American Cancer Society, Inc.

[2] M. Heath, K. W. Bowyer, D. Kopans, et al., “The Digital Database for Screening Mammography,” presented at 5th International Workshop on Digital Mammography (Toronto, Canada, 2000).

[3] University of South Florida Digital Mammography Home Page http://marathon.csee.usf.edu/Mammography/Database.html

[4] Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, Yu-Feng Li: Multi-instance multi-label learning. Artif. Intell. 176(1): 2291-2320 (2012)

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell., 24(5):603 C619, 2002.

[6] Spangler ML, Zuley ML, Sumkin JH et al Detection and classification of calcifications on digital breast tomosynthesis and 2D digital mammography: a comparison. AJR Am J Roentgenol 196:320–324 (2011)

[7] J. Bozek, M. Mustra, K. Delac, and M. Grgic, "A survey of image processing algorithms in digital mammography," J. Recent Advances in Multimedia Signal Processing and Communications, vol. 231, pp. 631-657, 2009.

[8] Jun Yang, MILL: A Multiple Instance Learning Library, http://www.cs.cmu.edu/~juny/MILL.

[9] Y.-F. Li, J. T. Kwok, I. W. Tsang, and Z.-H. Zhou. A convex method for locating regions of interest with multi-instance learning. In: Proceedings of the 20th European Conference on Machine Learning (ECML'09), Bled, Slovenia, 2009, pp.17-32.

[10] S. Andrews, I. Tsochantaridis, T. Hofmann. Support Vector Machines for Multiple-Instance Learning. NIPS 2002.

[11] A. Papadopoulos, D. I. Fotiadis, and A. Likas, “Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines,” Artif. Intell. Med. 34(2), 141–150 (2005).

[12] J. Yang, "Review of multi-instance learning and its application," 2001. [Online]. Available: http://www.cs.cmu.edu/juny/MILL/ review.htm