·生物电子学·

使用遗传算法的乳腺微钙化点特征优化

王瑞平[1,2]，万柏坤[2]，高上凯[3]

(1. 北京交通大学生物医学工程系 北京 海淀区 100044; 2. 天津大学生物医学工程系 天津 南开区 300072;
3. 清华大学生物医学工程系 北京 海淀区 100084)

【摘要】乳腺微钙化点包含众多属性，由于其中存在的冗余和不相关属性降低了微钙化点病变类型判别的性能。因此，特征子集选择问题成为微钙化点病变类型识别中的重要问题。该文针对传统优化方法用于特征选择的种种缺陷，提出了基于遗传算法的特征子集选择测算法。经乳腺微钙化点特征选择实例分析，证明该方法拥有较强的并行性和寻优能力，在特征选择领域有广阔的应用前景。

关 键 词 微钙化点; 特征子集; 遗传算法; 特征优化
中图分类号 TH776; TN911.73 文献标识码 A

Microcalcification Feature Selection in Mammograms Using Genetic Algorithm

WANG Rui-ping[1,2]，WAN Bai-kun[2]，GAO Shang-kai[3]

(1. Department of Biomedical Engineering, Beijing Jiaotong University Haidian Beijing 100044;
2. Department of Biomedical Engineering, Tianjin University Nankai Tianjin 300072;
3. Department of Biomedical Engineering, Tsinghua University Haidian Beijing 100084)

Abstract Microcalcifications include many redundant and irrelated features, which degrade the microcalcifications classification performance. So, feature subset selection becomes one of the important research issues in the process of microcalcification identification. In view of the deficiencies in traditional combination optimization method, an algorithm of feature subset selection based on genetic algorithm is proposed in this paper. According to the results of practical microcalcification classification example, it is proved that this method possess excellent parallelism and optimization performance.

Key words microcalcification; feature subset; genetic algorithm; feature optimization

乳腺钼靶X片上微钙化点病变类型的判别是一个典型的模式分类问题[1]。模式识别的任务是利用从样本中抽取出的特征将样本划分为相应的模式类别。特征向量中只有包含足够的类别信息，才有可能通过分类器完成无差错的模式分类。由于难以确定特征中是否已包含足够的类别信息，为了提高识别正确率，尽可能地增加特征数目以进行类别识别。但受分类器规模、训练过程的复杂性以及计算机容量等诸多因素的制约，过于庞大的特征维数往往不能取得良好的效果，因此需要采取措施，在不降低识别效果的前提下尽量减少特征维数。为了提高识别精度，在设计分类器前，必须去除两类冗余特征量：(1) 与分类目标无关的特征量；(2) 与其他特征量有较高相关性的特征量，即从一组数量为D的特征中选择出数量为$d(D>d)$的一组最优特征来，使得分类错误率最小。为此需要解决两个问题：(1) 选择的标准，即采用何种类别可分离判据；(2) 采用何种寻优方法来解决这一组合优化问题。围绕以上两个问题，本文采用类内－类间距离判据作为类别可分离判据，以遗传算法求解最优化的特征矢量。

1 类内–类间距离判据

从理论上讲，使分类错误概率最小的特征集应是最优的，但由于不易获得各类别条件概率分布密度，因而无法直接计算分类错误概率，需要更实用可计算的判据以衡量各类间的可分性，而该判据应

收稿日期：2005 - 03 - 10
基金项目：中国博士后科学基金资助项目(2004036063)
作者简介：王瑞平(1974 –)，女，副教授，主要从事生物医学信号和图像处理方面的研究.

当与错误概率有较好的单调关系。常用的可分性判据为基于信息论的特征–类(特征)互信息指标[2]和类内–类间距离判据[3]。前者是利用最大化特征-类信息去除第一类冗余特征，利用最小化特征–特征信息量去除第二类冗余特征。由于该算法必须对各个特征量进行离散化处理，需要有一定的先验知识，而且计算量较大，因此影响了它的使用效果，故本文选用类内–类间距离判据作为可分性准则。各类样本可以分开是因为它们位于特征空间的不同区域。如果两类样本之间的距离越小，而异类样本之间的距离越大，则分类效果越好。本文分别以类内散度矩阵 $\boldsymbol{S}_{\mathrm{w}}$ 的迹和类间散度矩阵 $\boldsymbol{S}_{\mathrm{b}}$ 的迹来度量以上两个距离，进而给出了类–内类间距离判据为：

$$J=\frac{t_r\boldsymbol{S}_{\mathrm{b}}}{t_r\boldsymbol{S}_{\mathrm{W}}} \tag{1}$$

$$\boldsymbol{S}_{\mathrm{b}}=\sum_{i=1}^{C}p_i(\boldsymbol{m}_i-m)(\boldsymbol{m}_i-m)^{\mathrm{T}}$$

$$\boldsymbol{m}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}x_k^{(i)},\ m=\sum_{i=1}^{C}p_i\boldsymbol{m}_i \tag{2}$$

$$\boldsymbol{S}_{\mathrm{w}}=\sum_{i=1}^{C}p_i\frac{1}{n_i}\sum_{k=1}^{n_i}(x_k^{(i)}-\boldsymbol{m}_i)(x_k^{(i)}-\boldsymbol{m}_i)^{\mathrm{T}} \tag{3}$$

式中 $\boldsymbol{m}_i$ 表示第i类样本集的均值向量；m表示所有各类的样本集总平均量；$x_k^{(i)}$，$x_l^{(j)}$ 分别为 w_i 类及 w_j 类中的D维特征向量；c为类别数；n_i 为 w_i 类中样本数；n_j 为 w_j 类中样本数；p_i、p_j 是相应类别的先验概率。

2 基于遗传算法的特征选择

目前乳腺微钙化点特征寻优方法主要包括可以求得最优解的穷举法、分枝定界法及可以求得次优解的顺序前进法、顺序后退法、晶格型搜索法和遗传算法[4]。穷举法和分枝定界法以大量的时间消耗来获得最优解，所以并不常用。顺序前进法、顺序后退法只是一个简单的串行搜索，将遗漏掉大量的特征组合。晶格型搜索是基于图论的确定型搜索模式，在搜索区域较大时，可以获得较为理想的次优解，但耗时依然较大。遗传算法[5-6]是通过模拟生物的进化过程中的繁殖、变异和自然选择来求解最优化问题。由于其具有良好的并行性、通用性及稳健性、已成为信息科学、计算机科学和人工智能等诸多学科所关注的焦点。本文采用基于遗传算法的优化方法来求得微钙化点特征矢量集的最优子集。遗传算法的基本过程如图1所示。

用于微钙化点特征矢量集最优子集提取的遗传算法流程与通用的遗传算法流程图和图1类似，令染色体位串长度为 m (m 为微钙化点特征矢量的数目)，其中每一位基因若取为1，则选择相应的特征；若取为0，则表示不选。由于大多数特征将被保留，在初始化种群时，每一位取1的概率为0.8。给定染色体 c 的适应度函数为类内-类间距离判据：$F_{\mathrm{fitness}}(c)=t_{\mathrm{r}}(\boldsymbol{S}_{\mathrm{b}}(c))/t_{\mathrm{r}}(\boldsymbol{S}_{\mathrm{w}}(c))$。$F_{\mathrm{fitness}}(c)$ 的大小代表了染色体 c 在遗传训练样本集上的分类能力。

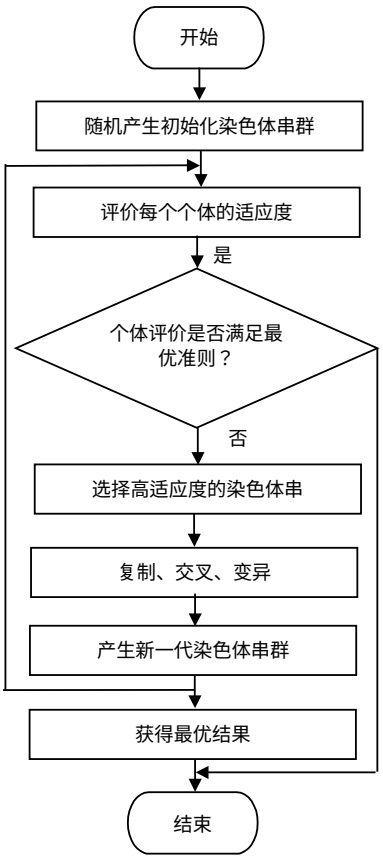

图1 遗传算法的一般流程图

采用的具体遗传算法如下：(1) 选择算子，对于每一个种群 Φ，先将上一代最优的染色体保留，其余染色体用轮盘赌的方法进行复制构成，即对于种群的染色体 c，其被选择的概率是 $P_{\Phi}=F_{\mathrm{fitness}}(c)/\sum_{c\in\Phi}F_{\mathrm{fitness}}(c)$；(2) 交叉算子，将染色体两两配对，然后随机的选择交叉点，将交叉点后的基因按交叉概率 P_c (本文取0.75为经验取值)进行互换；(3) 变异算子，对每个染色体按变异概率 P_m (本文取0.015为经验取值)改变基因的值，可以避免算法早熟。按照上述步骤进行迭代，直到每代中的最优染色体的适应度值保持一定的代数(本文选择20代)未改变，最后那些被选中的特征参数构成了最优特征矢量，用于下一步乳腺微钙化点的病变类型识别。

3　应用实例

本文选取了来自天津医科大学附属肿瘤医院100例乳腺感兴趣区域的173个恶性性微钙化点和165个良性微钙化点的形状、纹理、直方图等的33个特征参数[1]。将这些特征参数组成33维的特征矢量归一化到[0，1]范围后用遗传算法进行特征优化，最后得到几何特征参数圆度(F_3)、矩特征参数(F_5、F_9)、傅里叶描述子(F_{10})、与邻近微钙化点相关的特征参数(F_{17}、F_{20}、F_{23}、F_{24})和9个纹理特征的17个特征参数组成新的特征矢量用于下一步分类器的模式识别。图2所示为进化过程中最优个体的适应度变化趋势。从图中可看出，遗传算法具有较好的寻优性能。图3所示为进化过程中，最优个体染色体码的变化规律，不难看出，在迭代至186代时，最优个体包含了17个特征，相应特征集包含了圆度、傅里叶描述子、矩特征参数、纹理特征参数等。遗传算法选择以上特征参数与放射医学专家对乳腺钼靶图像上良恶性微钙化点在形状、纹理等方面特性的认识是相符合的。

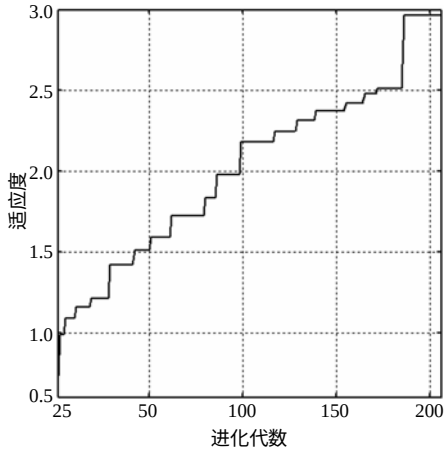

图2　进化过程中最优个体的适应度变化趋势

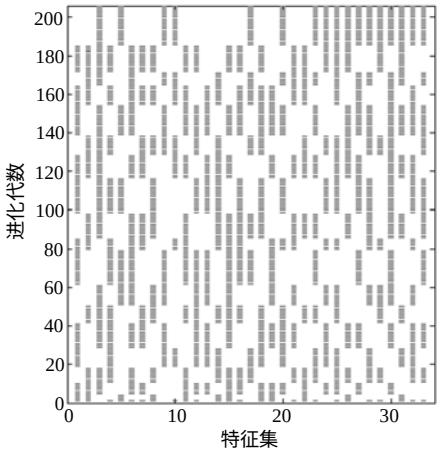

图3　最优个体在进化过程中的染色体码

本文用三层误差后向传播人工神经网络检验所选择特征参数表征良恶性微钙化点的能力和遗传算法的特征选择效果[1]。此处选取了来自100例乳腺感兴趣区域的173个恶性微钙化点和165个良性微钙化点组成人工神经网络的样本集，其中训练样本包括102个恶性微钙化点和100个良性微钙化点，检验样本包括71个恶性微钙化点和65个良性微钙化点。人工神经网络分类器经过训练表明，当用原始特征集进行微钙化点病变类型识别时，33-11-1的网络结构可以获得最佳分类性能(83.09%(113/136)，63个“1”和50个“0”被正确分类)；当用优化后的特征矢量进行微钙化点病变类型识别时，17-8-1的网络结构具有最佳分类性能(84.56%(115/136)，63个“1”和52个“0”被正确分类)。因此，当输入矢量为原始特征和优化特征时，分别由33-11-1BPNN和17-8-1BPNN构成的判别模型最优[1]。

为了评价遗传算法对原始特征矢量的优化能力，本文根据不同阈值下，分类器对应的阳性检出率(True Positive Rate,TPR)和假阳性率(False Positive Rate，FPR)制作了接受者操作特征曲线(Receiver Operation Characteristic Curve，ROC)，如图4所示。从图上可以直观地看出，17-8-1 BPNN分类器比较凸，曲线下的面积为0.881 4，而30-11-1BPNN分类器所对应的ROC曲线下面积为0.850 3。因此，前者具有更好的分类能力，即优化特征矢量和对应17-8-1BPNN分类器组成的判别模型对于微钙化点病变类型识别应该具有更高的诊断价值。

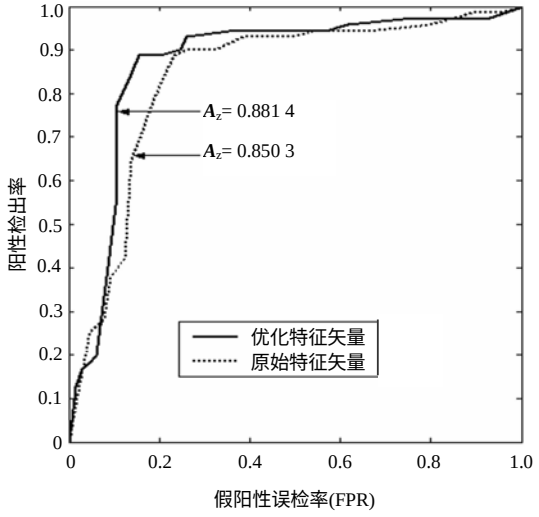

图4　微钙化点病变类型识别中的ROC曲线

(下转第153页)

的正逆表达式为：

$$\frac{\mathrm{d}^2\boldsymbol{q}}{\mathrm{d}t^2}=\boldsymbol{J}^{-1}\frac{\mathrm{d}^2\boldsymbol{p}}{\mathrm{d}t^2} \tag{17}$$

$$\frac{\mathrm{d}^2\boldsymbol{p}}{\mathrm{d}t^2}=\boldsymbol{J}\frac{\mathrm{d}^2\boldsymbol{q}}{\mathrm{d}t^2} \tag{18}$$

式中

$$\frac{\mathrm{d}^2\boldsymbol{p}}{\mathrm{d}t^2}=\left[\begin{array}{cccc}\dfrac{\mathrm{d}^2\Delta\alpha}{\mathrm{d}t^2} & \dfrac{\mathrm{d}^2\Delta\beta}{\mathrm{d}t^2} & \dfrac{\mathrm{d}^2\Delta p_Y}{\mathrm{d}t^2} & \dfrac{\mathrm{d}^2\Delta p_Z}{\mathrm{d}t^2}\end{array}\right]^{\mathrm{T}}$$

$$\frac{\mathrm{d}^2\boldsymbol{q}}{\mathrm{d}t^2}=\left[\begin{array}{cccc}\dfrac{\mathrm{d}^2\Delta\boldsymbol{S}_1}{\mathrm{d}t^2} & \dfrac{\mathrm{d}^2\Delta\boldsymbol{S}_2}{\mathrm{d}t^2} & \dfrac{\mathrm{d}^2\Delta\boldsymbol{S}_3}{\mathrm{d}t^2} & \dfrac{\mathrm{d}^2\Delta\boldsymbol{S}_4}{\mathrm{d}t^2}\end{array}\right]^{\mathrm{T}}$$

4　结　论

本文对2RPS+2TPS型微操作机器人进行了运动学分析，给出了推导微位移增量矩阵的一般方法。分析结果表明，微动机器人的特征矩阵均为常数矩阵。上述研究工作为微操作机器人的运动性能优化、动力学及控制的研究奠定了理论基础。

参　考　文　献

[1] CHEN Wen-jia, ZHAO Ming-yang, CHEN Shu-hong. A novel 4-DOF parallel manipulator and its kinematic modeling[C]// Proc. of IEEE, International Conference on Robotics and Automation, Seoul, Korea, 2001.

[2] LEE M K, PARK K W. Kinematic and dynamic analysis of a double parallel manipulator for enlarging workspace and avoiding singularities[J]. IEEE Transactions on Robotics and Automation, 1999, 15: 1025-1033.

[3] 李剑锋, 刘德忠, 潘新文. 3-PRS微操作器的运动学性质分析[J]. 中国机械工程, 2003, 14(14): 1194-1196.

[4] 孙立宁, 安　辉, 张　涛. 微动机器人运动学分析的基础研究[J]. 仪器仪表学报, 1998, 19(5): 464-471.

[5] 黄　真, 孔令富, 方跃法. 并联机器人机构学理论及控制[M]. 北京: 机械工业出版社, 1997.

编　辑　黄　莘

(上接第139页)

4　结　论

本文针对传统优化算法用于特征选择的不足，提出了以类内–类间距离为适合度函数，基于遗传算法的特征选择策略。该方法能够充分利用遗传算法的隐并行性，寻找最优的持征集合进行分类器设计。经微钙化点特征矢量集最优子集提取实例验证，该方法拥有强大的并行性和寻优能力，能高效地剔除原始特征集的冗余特征。将生成的优化特征集合用于神经网络分类器训练，能够提高微钙化点病变类型的识别精度。

本文研究工作得到北京交通大学科技基金(2005RC045)资助，在此表示感谢。

参　考　文　献

[1] 王瑞平. 乳腺X线影像的计算机辅助诊断新方法研究[D]. 天津: 天津大学, 2003.

[2] NOJUN K, CHOI Chong-Ho. Input feature selection by mutual information based on parzen window[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1667-1671.

[3] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000: 178-179.

[4] 王　凌. 智能优化算法及其应用[M]. 北京: 清华大学出版社, 2001: 1-14.

[5] GOLDBERG D E. Genetic algorithms in search, optimization and machine learning[M]. New York: Addison Weseley, 1989: 165-176.

[6]GABOR R, ANIKO E. Genetic algorithms in computer aided design[J]. Computer-Aided Design, 2003, 35(8): 709-726.

编　辑　孙晓丹