第 8 卷 第 3 期 自 动 化 学 报 Vol. 8, No. 3
1982 年 7 月 ACTA AUTOMATICA SINICA July 1982

# 宫颈上皮细胞图象的计算机分类

陈传涓 刘树范[1)]
（中国科学院生物物理所） （医学科学院肿瘤研究所）

## 摘 要

通过对约 300 个宫颈上皮细胞图象的分析研究，结果表明：(1)当将对象分为三类（正常、核异质与癌）时，三级线性判决的效果明显地优于一级判决；(2)从纯数学角度选取的非视觉特征对分类是重要的．在适当地选择训练集且限定每次判决所用特征数为五时，正确识别率可达 75%．

## 一、引 言

细胞图象自动分析是模式识别与生物医学工程当前的研究重点之一．作者以往的研究[1]表明，对于食道细胞的两类判决问题，线性判决可以得到较好的结果．

对宫颈上皮细胞图象进行三类（正常、核异质与癌）判决的多次实验[2,3]结果表明，由于样本分布形状的限制，一次判决不可能取得理想的效果．例如，当利用双重回归[4]构造线性判决函数时，即使在训练集上正确识别率也仅能达到 50% 左右．因此，作者采用三级线性判决树[5]进行三类判别．

在判决树的每一级上必须进行特征选择．从实际需要出发，限定每次判决所用特征数为 3—5 个．在特征选择时采用了比较单个特征分类能力和双重筛选逐步回归两种方案．实验结果表明两种方法互有优劣．

为提高分类能力，将所取的视觉特征从原有的十八个增加到二十个，并且引入了八个非视觉特征．特征选择的结果说明，新增加的非视觉特征对于分类具有一定的意义．

## 二、系统与流程

细胞图象分析系统由 Scanning Microscopemeter-05（西德，Opton）和国产 TQ-16 电子计算机组成．前者用于细胞图象数据的获取，结果以穿孔纸带形式输出．后者则用于其它环节的处理．系统的流程如图 1 所示．

数据的获取包括取得样品、涂片制备和模-数转换等步骤[1]．本文将主要讨论特征抽提、分类器设计和特征选择等部分．

本文于 1981 年 2 月 17 日收到．

1) 周彬、赵宁二同志参加了本项研究的技术工作．

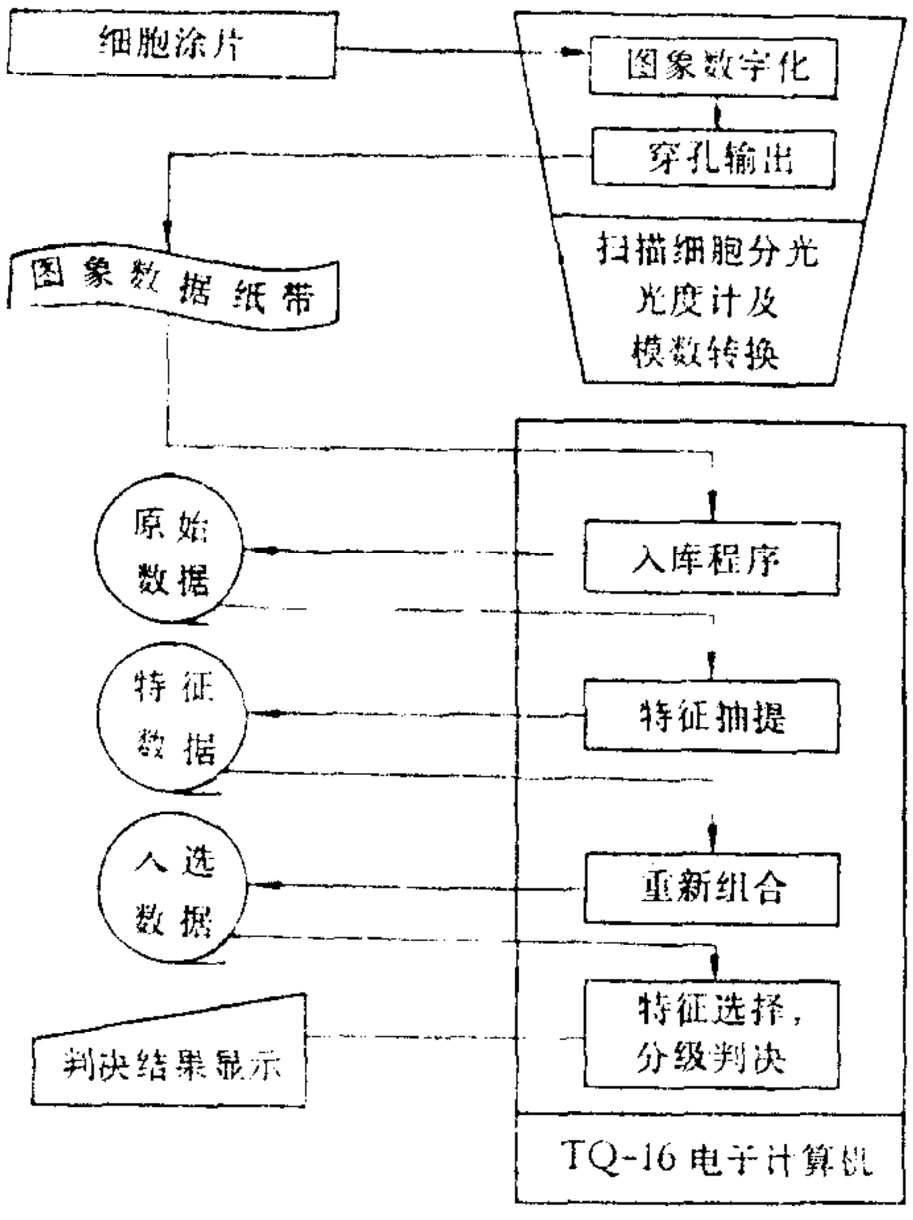


图 1 细胞图象分析系统

## 三、特征抽提与分类器结构

### 1. 非视觉特征的引入

所谓视觉特征通常应满足以下条件：① 可以指出其与细胞形态或结构的关系；② 数值依赖于图象分割结果. 当分割阈值改变时，细胞核与胞浆部分的大小与形态将随之发生变化，而视觉特征的数值也将改变.

考虑到在特征抽提过程中从纯数学角度引入的非视觉特征已日益受到重视，在原有十八个特征[2]的基础上引入了八个直方图特征，记作特征 21—28[1]. 将一细胞图象内可能出现的光密度值划分为八个灰阶范围，用 $F_{21}^{*}$—$F_{28}^{*}$ 分别表示一细胞图象中属于范围 1—8 的点数，再各自除以该图象内的总点数即得特征 21—28.

以上八个特征属于非视觉特征，并不能对其做出直观解释. 但是，对于描述那些结构十分复杂，在分割中难以准确求出细胞核阈值的图象，上述特征将起重要的作用.

### 2. 多级线性判决结构

在评价一个实用的模式识别系统时，分类效果和工作效率是两个重要的标准. 对本文所讨论的三类问题，一级线性判决固然计算量小，但效果很差，训练集上的误识率(仅是实际误识率的下界[6])已达 50% 左右. 为使系统可以用于实践，采用三级线性判决结构. 这种结构不是唯一确定的. 经过对多种结构的比较研究，表明图 2 所示的结构效果较好. 对同一批数据所做的多维尺度变换和二维显示[7]的结果也证实了这一点.

图 2 所示的结构是一个二元树. 每个结点处对应一种分类状态和一个线性判决函数

1) 新引入的特征 19—20 分别为胞核内各点均方差和光密度平方和，仍属于视觉特征.

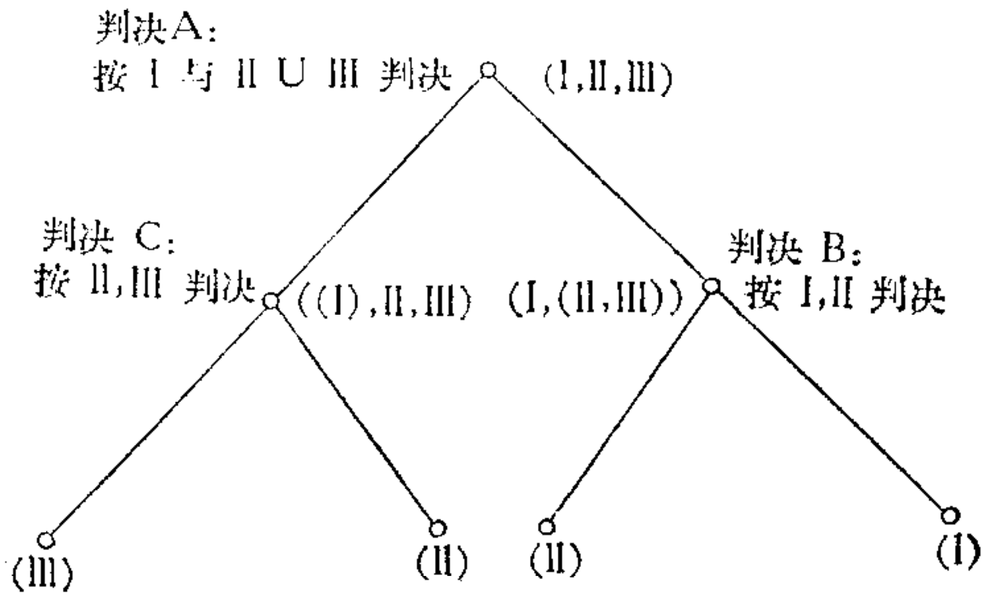


图 2　三级线性判决结构

(叶子除外). 在图中用 I, II 和 III 分别表示正常、核异质和癌三类. 树根对应于三类混合状态. 由于 II, III 两类混淆较为严重，所以在第一级上首先按 I 和 (II∪III) 予以分类判决，使判决界线偏于后者. 在第二级上，对右结点按 I,II 两类判决，其中混入的 III 类点不予剔除，在第三级上将其判为误识样品；对左结点按 II, III 两类判决，其中混入的 I 类点判为拒识，并在下一级判决中剔除.

由于采用二元树结构，所以在每个结点上进行特征选择时可按两类问题进行，从而使问题简化.

## 四、特征选择和训练集选取

### 1. 特征选择

为提高判决效率和排除特征过多带来的不利影响，在每步判决中只取 3—5 个特征. 特征选择包括以下两种方案:

1）考查单个特征分类能力

对每个特征的分类能力用以下公式测度:

$$W_i = |M_i^{(1)} - M_i^{(2)}|/(\sigma_i^{(1)} + \sigma_i^{(2)}). \qquad (1)$$

其中 $M_i^{(1)}$, $M_i^{(2)}$ 分别为第 1,2 类样品中第 $i$ 特征均值，$\sigma_i^{(1)}$, $\sigma_i^{(2)}$ 分别为第 1, 2 类样品中第 $i$ 特征均方差. 由于 $W_i$ 为无量纲量，所以不必对原始数据标准化.

在判决树的每个结点上需要分别进行特征选择，步骤为: ① 计算全部特征的相关系数矩阵. 若某两特征间的相关系数大于某阈值，则在今后各步中只取其中之一，以充分发挥各入选特征经组合后的分类效果. ② 对被保留诸特征分别按式 (1) 计算 $W_i$ 值，并依 $W_i$ 大小将各特征排序. ③根据要求取前 $k$ 个特征，$k$ 为限定的特征数目. 在每个结点上选出了所需的 $k$ 个特征后，用 Fisher 方法[7]构造判决函数.

上述方法的优点为: ① 由于对各特征分别考虑，所以当引入新的特征时只需单独计算其 $W_i$ 值并排序，而不必改变以前的计算结果. 这对正处于不断完善过程中的系统显然是可取的;②由于只需根据对 $W_i$ 的排序选取特征，因此易于控制所选特征的个数.

2）双重筛选逐步回归方法

已经证明，在两类情形下本方法等价于逐步判别分析方法. 本方法的优点为: ① 误识率稍低于前一方法;②每当选入一特征时已充分考虑到它与已入选特征的相关性质，从而弥补了前一方法的不足. 但是，由于本方法的入选特征数需由 $F$ 检验参量确定，因此当

两类样品严重混淆时往往难以使人选特征数目恰好与预期的的个数一致.

**2. 样品及训练集选取**

本文所述实验由于设备等条件所限，图象分割全部由计算机自动进行[1]而未通过人-机会话加以修正，而且未及引入形态特征[8]，只是按自动图象分割结果与照片吻合程度较高、形状一般，没有特殊的状态畸变的原则，从 550 个宫颈上皮细胞中选取了 279 个用作训练和考试样品.

全部样品分为两组. 第 54 组共 107 个样品（I 29, II 3, III 75）作为考试集. 第 60 组共 172 个样品（I 59, II 60, III 53），按以下两种方式进行组合以形成训练和考试集：① 进行随机排列后分别取前 75 个和后 75 个，而形成第 63 组（I 22, II 23, III 30）和第 64 组（I 20, II 34, III 21）. ② 由细胞病理学家先将本组样品分为三类，再将每类中编号为奇数及偶数的样品分别组成第 66 和 67 组，两组各含 85 个样品（I29, II 30, III 26）.

## 五、实 验 结 果

实验采用第 66 组样品为训练集，第 67 和 54 组为考试集. 在判决树的每个结点上分别取 3 个和 5 个特征. 计算结果分别见表 1 和表 2 .

表 1 取三个特征时的分类结果

| 特征选择方案 | | 计算单个特征分类能力 | | | | | 双重筛选逐步回归 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 各次入选特征号数 | 1 | 15, 2, 27 | | | | | 15, 7, 20 | | | | |
| | 2 | 15, 7, 2 | | | | | 15, 7, 14 | | | | |
| | 3 | 15, 7, 2 | | | | | 15, 5, 27 | | | | |
| 识别结果与正确识别率 | 66 组（训练集） | 人＼机 | 1 | 2 | 3 | 拒识 | 人＼机 | 1 | 2 | 3 | 拒识 |
| | | 1 | 25 | 3 | | 1 | 1 | 26 | 2 | | 1 |
| | | 2 | 2 | 16 | 12 | | 2 | 1 | 17 | 12 | |
| | | 3 | 1 | 3 | 22 | | 3 | | 1 | 25 | |
| | | 74.11% | | | | | 80.00% | | | | |
| | 67 组 | 人＼机 | 1 | 2 | 3 | 拒识 | 人＼机 | 1 | 2 | 3 | 拒识 |
| | | 1 | 23 | 5 | | 1 | 1 | 25 | 2 | | 2 |
| | | 2 | 3 | 16 | 11 | | 2 | 6 | 8 | 16 | |
| | | 3 | | 4 | 22 | | 3 | | 1 | 25 | |
| | | 71.76% | | | | | 68.23% | | | | |
| | 54 组 | 人＼机 | 1 | 2 | 3 | 拒识 | 人＼机 | 1 | 2 | 3 | 拒识 |
| | | 1 | 29 | | | | 1 | 29 | | | |
| | | 2 | 2 | 1 | | | 2 | 2 | 1 | | |
| | | 3 | 1 | 23 | 51 | | 3 | 2 | 20 | 53 | |
| | | 75.70% | | | | | 77.57% | | | | |

表 2　取五个特征时的分类结果

<table>
<tr><td colspan="2">特征选择方案</td><td>计算单个特征分类能力</td><td>双重筛选逐步回归</td></tr>
<tr><td rowspan="12">识别结果与正确识别率</td><td>每次入选特征号数 1</td><td>15, 2, 27, 28, 26</td><td>15, 7, 20, 18, 21</td></tr>
<tr><td>每次入选特征号数 2</td><td>15, 7, 2, 27, 6</td><td>15, 7, 14, 26, 3</td></tr>
<tr><td>每次入选特征号数 3</td><td>15, 7, 2, 17, 5</td><td>15, 2, 12, 5, 11</td></tr>
<tr><td rowspan="5">66 组</td><td>机\人 / 1 / 2 / 3 / 拒识<br>1 / 26 / 2 / / 1<br>2 / 1 / 13 / 13 /<br>3 / / / 26 /</td><td>机\人 / 1 / 2 / 3 / 拒识<br>1 / 26 / 2 / / 1<br>2 / 2 / 16 / 12 /<br>3 / / / 26 /</td></tr>
<tr><td colspan="2">76.47%</td><td>80.00%</td></tr>
<tr><td rowspan="5">67 组</td><td>机\人 / 1 / 2 / 3 / 拒识<br>1 / 22 / 4 / / 3<br>2 / 4 / 11 / 15 /<br>3 / / 3 / 23 /</td><td>机\人 / 1 / 2 / 3 / 拒识<br>1 / 24 / 4 / / 1<br>2 / 3 / 13 / 14 /<br>3 / / 2 / 24 /</td></tr>
<tr><td colspan="2">65.88%</td><td>71.16%</td></tr>
<tr><td rowspan="5">54 组</td><td>机\人 / 1 / 2 / 3 / 拒识<br>1 / 27 / 2 / / <br>2 / 1 / 2 / / <br>3 / / 26 / 49 /</td><td>机\人 / 1 / 2 / 3 / 拒识<br>1 / 27 / 2 / / <br>2 / / 3 / / <br>3 / 3 / 22 / 50 /</td></tr>
<tr><td colspan="2">72.89%</td><td>74.76%</td></tr>
</table>

当采用第 63 组为训练集，64 和 54 组为考试集时，训练集和考试集上的识别率相差 20% 以上. 这一现象是由于 63 组诸样品主要采自少数病例所致. 这表明对训练集的不同选择将影响判决效果.

另外，采用逐步回归方法时，选入特征的个数与入选阈值 $F$ 有关. 由于设备限制，对 $F$ 仅能取 10 位有效数字. 对本批样品，当取 $F=1.283892196$ 时，选入 8 个特征，取 $F=1.283892197$ 时则仅选入 2 个特征. 因此，采用逐步回归方法在通常的设备条件下不能保证恰好选入所需数目的特征. 表中所列出的特征为按入选顺序所取的前三个或前五个，判决系数则用 Fisher 方法重新求得.

综合上述结果可得表 3 所示的结论. 结论表明，逐步回归方法所得结果的识别率略高于单个选择方法. 但后者仍不失为一种可取的方案. 此外，结果表明增加特征个数

表 3　特征选择方案综合比较

| 识别率 / 方案 / 特征数 | 单个挑选法 | 逐步回归法 |
|---|---|---|
| 3 | 74.00% | 75.45% |
| 5 | 71.84% | 75.45% |

对于改善分类效果并无作用.

## 六、讨 论

根据上述结果,就几个问题讨论如下:

**1. 细胞图象自动分析研究的目的**

在目前情况下,研究的目的可以是: ① 研制具有高分辨率的多类分析系统. 本文所述工作是在这一方面的一次尝试. 但为了进一步改善结果,必须大量增加样品和特征数目,而这一目标的实现又必须依靠相应的设备条件; ②主要研制两类分析系统,而将努力方向集中在提高精确度上,即用定量方法去鉴别用经验或直接观察所难以确诊的问题. 根据作者的实践,这一目标可能更具有现实性.

**2. 非视觉特征引入的必要性**

本项研究中所用的550个细胞经分割后,由于与照片吻合程度不高而淘汰了大约40%. 这一现象说明,仅用细胞病理学者所提供的,来自经验的启发式视觉特征,是不足以充分描述细胞图象的. 虽然通过改进分割等处理方法可以改善上述结果,但对于一些核结构异常复杂的细胞,恐怕任何分割方式都难以克服与实际形状不符的现象. 根据对于一些自动分割效果较差的细胞所进行的分析,发现直方图特征对于描述这类细胞具有一定作用. 如前所述,这种非视觉特征在各级判决中都有重要意义. 因此可以设想,在继续引入各种非视觉特征后,将会在较大程度上弥补视觉特征的不足,甚至取代视觉特征的作用.

**3. 特征选择与训练集选取**

如前所述,实验采取的两种特征的选择方法互有优劣. 逐步回归方法虽然识别率稍高,但仍存在着计算量大等问题. 因此,仍有必要继续探讨更为理想的特征选择方案.

选取训练集的可靠方案仍有待进一步研究. 这类方案必须遵循的原则包括随机分布、足够多的样品数以及充分的代表性等等. 为选取恰当的训练集所设计的若干算法将在另文中讨论.

## 参 考 文 献


[1] 癌细胞自动识别研究协作组,细胞图象自动识别研究专栏,生物化学与生物物理进展,(1980),第3期.

[2] 蒋代梅、陈传涓等, 基于拟合优度的两个分类算法及其在细胞图象自动分析中的应用, 北京工业大学学报,(1982),第1期.

[3] 钟友良、陈传涓等,一种自动选择参数的聚类算法,北京工业大学学报,(1982),第1期.

[4] 张尧庭等,气象资料的统计分析方法,农业出版社,(1979),61—67.

[5] J. Taylor, et al., Automated Hierarchic Decision Structures for Moltiple Category Cell Classification by TICAS *ACTA Cytologica*, **22** (1978), No. 4, 261.

[6] Kanal L., Pattern in Pattern Recognition, *IEEE Trans. IT* **20** (1974), No. 6.

[7] 福永圭之介,统计图形识别导论(陶笃纯译),科学出版社(1978),382.

[8] J. J. Sychra, Cytoplasmic and Nuclear Shape Analysis for Computerized Cell Recognition, *ACTA Cytologica*, **20** (1976), No. 1.

# THE COMPUTERIZED SORTING FOR CERVICAL EPITHELIUMS

CHEN CHUANJUAN

*(Institute of Biophysics, Academia Sinica)*

LIU SHUFAN

*(Cancer Institute, Chineses Academy of Medical Sciences)*

**ABSTRACT**

About 300 cervical epithelium images were analysed and studied as a whole. The results have shown that 1) when these images are devided into three classes (normal, dysplastic, cancer), tertiary linear decision effect is much better than the primary level decision and 2) non-visual features that is acquired from pure-mathematical method are important for sorting.

When numbers of feature to each decision is restricted to five and training sets are preperly selected, the right recognition ratio achieved is 75%.