往落到 H_i 较大的属于核内某些深灰度级上，这是因为核内个别颗粒的存在引起灰度跳变大，介于其中的许多灰度出现的频数为零，虽经平滑处理，但频数的最低谷也常落到此灰度段中。这是造成方法(六)出现较大的负的 $\varepsilon_{SN}\%$ 的主要原因。若能将方法(六)找频数谷的方法限在胞浆和胞核两大频数包峰之间，或能先验地给出核阈值选择的范围以及核面积的下限，估计会大大改善这种方法的效果。

由于以上各种因素，细胞图象其各区域的灰度分布基本上是非等灰度分布，使得用"阈值法"划分各部分区域受到限制，尤其核阈值 Z_N 的选择带来较大的误差，我们第一阶段的试验结果不够理想。但阈值法的优点是比较简单，因此今后我们拟继续研究，以提高其准确性，为细胞的诊断和分类等后续环节作好准备。

参 考 文 献

[1] Wied, G. L. et al.: *Acta Cytol.*, **12**, (2) 180—204, 1968.

[2] Tanaka, N. (田中昇) et al.: *Acta. Cytol.*, **21**,(1) 85—89, 1977.

食管上皮细胞自动分类过程中的特征选择

——食管癌细胞自动分类的研究专题之四

陈传涓　楼 恒

(中国科学院生物物理研究所)

为满足计算机分析细胞的特殊要求，采用了前文介绍的细胞涂片制备方法。通过数据输入，细胞数字图象的预处理和区域划分等步骤，已将待分析的单个细胞的胞浆和胞核分别提取出来。在此基础上可以根据临床细胞学家的经验和广泛的数学模型二者尽可能多地抽取胞浆与胞核的形状与结构特征。

究竟哪些特征对鉴别诊断或判决分类有意义？如何从大量已抽取的特征中进行选择？这一特征选择问题在模式识别的概率统计方法中是一个重要组成部分，例如美国 TICAS (Taxonomic Intra-Cellular Analytic System) 系统程序包内的子程序 DSELECT 子程序即专司特征选择的功能[1,2]，日本 CYBEST (Cyto-Biological Electronic Screener by Toshiba) 系统基础研究的大量内容也涉及特征选择[3]。

本文采用了依赖于 F 统计量临界值选取的多自变量及多因变量双重筛选逐步回归方法。该方法在选择对分类判决最有效的特征的同时给出由这些特征线性组合而成的判决函数。在数据获得方式尚未完善和特征抽提不够广泛的情况下，对所用训练样品得到了较为满意的分类效果。

一、材料与方法

细胞原始数据均用 OPTON SMP 光学扫描显微镜记录，空间分辨率分别为 4.0 微米和 1.0 微米，光度分辨率为 7 比特，波长 500 毫微米。被记录的数据以纸带形式经快速穿孔输出并转入国产 013 计算机原始数据库。经过特征抽提子程序对原始数据库中各细胞数据进行信息压缩后，再存入特征数据库。本项工作所用特征数据直接引自特征数据库。

目前共使用 18 种特征，每种特征含义如下：

特征 1：NA，细胞核面积，以扫描点个数为单位。

特征 2：NA%，胞核面积在整个细胞面积内所占百分比 (NA/NA + CA)。

特征 3：NTE，核内各点光密度总和(总光

密度)。

特征 4：NTE%，核内光密度和在整个细胞总光密度中所占百分比。

特征 5：NME，核平均光密度（NTE/NA)。

特征 6：NTI 核内各点透过率总和。

特征 7：NTI%，核的透过率总和在整个细胞透过率总和中所占百分比。

特征 8：NMT，核平均透过率（NTI/NA)。

特征 9—16：分别为胞浆的有关特征，含义同胞核。如特征 9：CA，胞浆面积。又如特征 16：CMT，胞浆平均透过率。

特征 17：NDE，胞核范围内的分布误差 NME — log NMI/NME，反映核的不均匀度。

特征 18：胞核范围内同一行中相邻两点透过率之差的总平均值 $\Sigma|T_i - T_{i-1}|/NA$，其中 T_i，T_{i-1} 分别为属于同一行的两邻点的透过率。

以上特征主要根据临床病理学家的诊断经验并参照 APMOS-II 的某些分析指标选择。

不难看出，上述诸特征的选择均依赖于对显微图象扫描数据的分割，即区分代表胞核与代表胞浆的数据。

还可看出上述诸特征间相关性很大。由于特征选择时采用线性回归方法，所以，当某一特征被选出后，与它具有线性相关的各个特征在选择过程中将被淘汰，而其他非线性相关特征仍有其独立意义。

依赖于 F 统计量临界值选取的双重筛选逐步回归方法[4]通常在多元统计分析中用以处理多个自变量与多个因变量间的线性回归问题。此方法较为接近细胞病理学家根据不同特征确定细胞类型的认识过程。它以相关性大小为判据而选择特征并采用反复剔除过程，即每当新入选一特征后就对全部已入选特征重新筛选，并淘汰可由新入选特征取代的已入选特征。这一过程对于选择关系密切而非线性相关的各种特征至为必要，这一点已由我们的工作证实。

临界值 F 的选取原则为(一)使入选特征数在 3—7 之间及(二)误识率最小。当 F 值减少时，入选特征数将随之递增，但识别率未必随之改善，即使对训练集本身也是如此。对于大量的试验样品，训练样品数量不是非常大时，用于判决分类的特征数增加将导致分类效果的变坏。因此，对特征数的限制是必要的。

在数学上，双重筛选方法是在回归的残差平方和最小的意义下进行的，而不直接涉及识别率问题。当将本方法用于逐步判别(例如，对两类问题，分别令两类训练样品的因变量为 1 和 0)时，可将求得的回归系数作为线性判决函数，但此结果并不保证误识率最小。在我们的计算程序中增加了对上述方法所得结果进行误识率检验的内容，并根据误识率最小的标准对一系列 F 值及入选特征进行评价，从中最后选出最理想的结果。

结　　果

训练样品取自一段时期内记录的细胞数据，原始分类由有经验的医生作出。须指出，由于不同时刻仪器状态及细胞染色等差别，会引起一定误差。大批量处理时应该随机抽取样品。由于目前记录的细胞数据不多，所以暂时采用上述权宜的处理。基于相同的原因，暂时采用在训练集上进行检验的方法(重复检验)。

用两种方法获得原始数据。其一是用 4 微米步距扫描，对象为各层正常与不正常细胞；其二是用 1 微米步距扫描，对象为重增以上细胞。

4 微米数据共 97 个。其中正常（包括轻增）35 个，异常(重增 I，重增 II，癌) 62 个。第一步工作暂作为两类问题而未作分层处理。1 微米数据 56 个。其中重增 I 12 个，重增 II 及癌 44 个。在 013 计算机上的计算结果如表 1 及表 2 所示。

以识别率最高为标准时，4 微米粗扫描数据以选取特征 4(NTE%)，6(NTI)，2(NA%)，7(NTI%) 和 8(NMT) 为最佳。用上述五特征从线性回归残差平方和最小出发构成线性判决函数（系数依次为 -0.3928787×10^{-1}，0.4888622×10^{-3}，0.6463819×10^{-1}，-0.4162628

表 1　4 微米数据计算结果

F	入选特征		误识细胞		
	个数	编号(按入选顺序排列)	个数	编号	误识率
4	3	4, 6, 2	9	1, 24, 40, 61, 73, 79, 85, 90, 91	9.3%
3, 2	5	4, 6, 2, 8, 7	7	1, 24, 40, 73, 79, 85, 91,	7.2%
1, 0.8, 0.6, 0.4	10	4, 6, 2, 8, 7, 13, 16, 18, 17, 5	7	同 F = 3	7.2%
0.2	11	4, 6, 2, 8, 7, 13, 16, 18, 17, 5, 1	9	1, 24, 39, 40, 42, 73, 79, 85, 91	9.3%

表 2　1 微米数据计算结果

F	入选特征*		误识细胞		
	个数	编号(按入选顺序排列)	个数	编号	误识率
4.3	1	3	19	11, 13, 14, 16, 17, 21, 22, 25, 26, 27, 28, 33, 34, 38, 39, 40, 41, 42, 49	33.9%
2	2	3, 14, 8, 15, −3, −14	18	8, 11, 12, 13, 15, 16, 17, 22, 23, 25, 28, 33, 37, 38, 39, 41, 54, 55	32.1%
1	4	3, 14, 8, 15, −3, −14, 5, 11	14	11, 12, 13, 15, 16, 17, 22, 23, 25, 28, 33, 38, 39, 41	25.0%
0.8	4	3, 14, 8, 15, −3, 17	13	12, 15, 16, 17, 22, 25, 28, 33, 37, 38, 39, 41, 55	23.2%
0.6	5	3, 14, 8, 15, −3, 17, 5	14	8, 12, 15, 16, 17, 22, 25, 28, 33, 38, 39, 41, 55, 56	25.0%
0.4	5	3, 14, 8, 15, −3, 17, 5, 9, −14	10	15, 16, 17, 22, 25, 28, 33, 38, 39, 41,	17.9%
0.2	8	3, 14, 8, 15, −3, 17, 5, 9, −14, 6, 14, 11	11	13, 15, 16, 17, 22, 25, 28, 33, 39, 41, 56	19.6%

* 负号表示入选后又剔除的特征。

$\times 10^{-1}$, 0.1090564×10^{-1}), 映射到一维空间进行判决时,在训练集上的识别率为 92.8%。

对 1 微米数据的最佳选择方式为选取特征 8(NMT), 15(CTI%), 17(NDE), 5(NME) 和 9(CA)。线性判决函数系数依次为 0.4109839×10^{-1}, 0.1533177×10^{-1}, 0.3873024×10^{-1}, 0.1715812×10^{-1} 和 $-0.7522685 \times 10^{-1}$。对训练集的识别率为 82.1%。

讨　论

1. 结果表明 1 微米扫描数据识别率反而低于 4 微米数据。这是由于后者分类要求粗糙,只须在正常与异常二者间作出判决。而前者则要求对重增 I 和重增 II 以上的细胞分类,因而难度较大。从入选特征同样可见,对于后者,只要利用核、浆面积及颜色深浅即可得到较高的识别率;而前者即使加入核均匀度等特征,识别效果仍不及后者。这一现象也与初学细胞分类的技术员的实际经验吻合。因此,脱离分类要求去谈论分类精度与空间分辨率的关系似难得到一般结论。由此也易于理解何以日本 CYBEST 系统采用 4 微米扫描,而美国 Wied 等人则倾向于否定 4 微米分辨率[5]。显然,这是由于各自的分类要求不同所致。

2. 在两类数据中,无论选择何种特征,被误识细胞都具有较大重复性。现对在每种特征组

合下均被误识的细胞分析说明如下：1)对于食管上皮各层细胞，4 微米数据仅分为正常和异常两类，因而误分不可避免。

按照细胞病理学家的标准，食道上皮正常细胞从底层到表层，胞核直径由 8—10 微米减少到 4 微米，核浆比由 1:2—3 下降到 1:8—12（图 1 见封三）。

由于试验样品系经过食管拉网获得，所以训练样品中多数正常细胞属于表浅层，而多数癌细胞将以底层小圆癌细胞为主。因为只分两类，所以一旦出现底层或中深层的正常细胞（如 1 号、24 号细胞，见图 2 见封 3），则将出现一类错误；而当出现中，表层重增或退化细胞（如 40 号，见图 3 见封 3）时将出现二类错误。

被误判的79号细胞数据获得失真，85 号则是一裸核细胞，计算机自动找胞核时将核仁误认为胞核而引起误判。

上述分析表明，为提高系统识别率，应首先改进数据获得方式，使数据与显微图形的深浅尽可能一致。此外还应采用先分层，再分类的分级判决机构，增加分类类别。日本、美国的细胞分析工作近来也逐渐趋向所谓判决树方向[6,7]。

2）为正确区分底层重增 I 与重增 II 以上细胞，应考虑改进特征抽提的数学模型和引入新的特征。

实践表明，1 微米数据判决中多数错误属于第二类，即将重增 II 以上细胞误判为重增 I，而误判原因主要是特征不够。例如，细胞病理学家判断 15、38 号细胞的癌时，细胞形态是一重要指标。又如，第 16、17、33 与 41 号细胞的核位置偏心，紧贴细胞膜，这也是一个重要指标（图 4 见封三）。但是，上述指标均包含在所用 18 个特征之内。再如，关于核结构的均匀度是医学专家最常用的术语之一。特征 17、18 虽可部分地反映这一指标（而且，在特征选择中特征 17 也已入选），但仔细分析，只用这两个特征仍感不足。

3. 最后，对于模式识别中常用的其他特征选择方法，诸如特征组合的最小熵方法，逐个淘汰特征方法，Kruskal-Wallis 检验等，尚需进一步分析对比，以判断它们用于细胞识别时的效果与特点。

参 考 文 献

[1] Bartels, P. H. et al.: *Acta Cytol.*, **14**, 486—494, 1970.

[2] Bartels, P. H. et al.: *Appl. Opt.*, **9**, 2453—2458, 1970.

[3] Y. Imasato et al.: *Computer in Biology & Medicine*, **5**(3): 245, 1975.

[4] 张尧庭，方开泰：《多元分析》，科学出版社（待出版）。

[5] Bartels, P. H. et al.: *Acta Cytol.*, **21**, 753—764, 1977.

[6] Taylor, J. et al.: *Acta Cytol.*, **18**, 512—521, 1974.

[7] 铃木等：电子通信学会パタン认识と学习专门委员会资料，1976，9 月。

食管上皮细胞分类判决方法的研究

——食管癌细胞自动分类研究专题之五

阎 平 凡

（清华大学自动化系）

一、方 法 概 述

研究食管上皮细胞的识别分类问题，除如前文所介绍的在双重筛选过程中加以判决外，我们还采用了 Fisher 方法[1]，因为它有以下优点：

1. 对于只分成两类且每类分布密度是多维正态时，它与 Bayes 判决一样能给出最优判决。

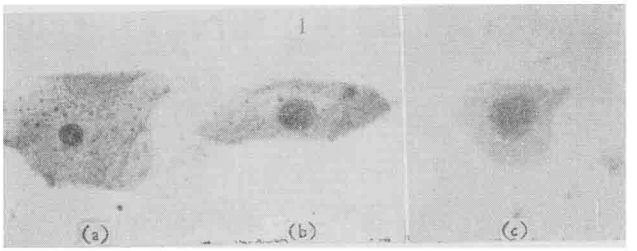

图 1　食管上皮表浅 (a)、中 (b)、深 (c) 层正常细胞照片

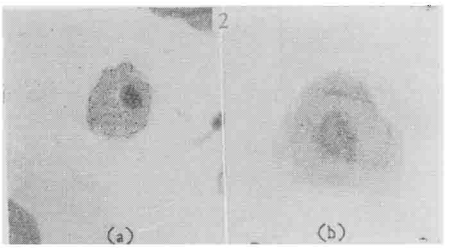

图 2　被误判为不正常细胞的中 (a)、深 (b) 层正常细胞

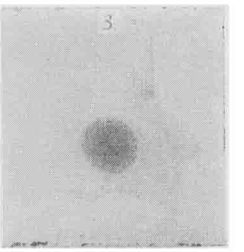

图 3　被误判的 40 号细胞

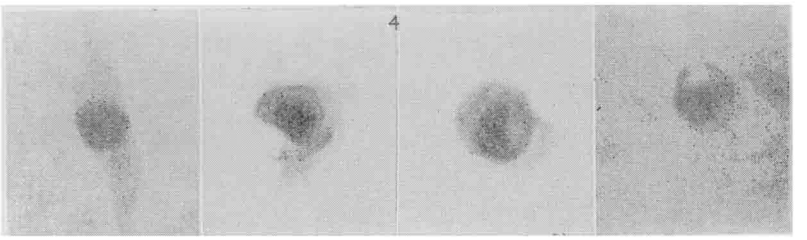

图 4　误判细胞的几种特殊形态

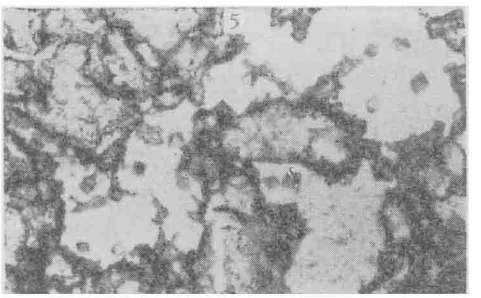

图 5　组织铁蛋白晶体示例(肝)