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。

实验结果表明，食管上皮细胞各个特征的分布虽然不能认为完全是正态的，但分布规律比较简单，例如基本都是单峰的，有的还比较对称，用 Fisher 判别法可以得到较好的结果。

2. 对细胞分类，两种错误的损失是不一样的，有时要求假阴性率不能超过某一给定值，用 Fisher 方法投影成一维分布后调整判决阈值比较简单，相比之下线性判决方法或其他非参数法则有一定困难。

3. 方法简单，计算工作量较小。

Fisher 方法的基本出发点是把多维问题投影成一维的，投影方向的选择应使两类分得最开，具体说就是把原来多个特征经线性组合后化为一个新的特征，选择组合式中各原始特征的系数，使在新的特征下类内离差小而类间离差大，经过推导证明求系数的过程是解一个线性方程组，因此可以利用现有的解线性方程组的程序在计算机上求解。

二、实 验 结 果

对两组数据进行了分析：

第一组：扫描步距 4μ，共 97 个细胞，其中 35 个正常，62 个异常。

第二组：扫描步距 1μ，共 56 个细胞，其中 12 个正常，44 个异常。

特征选取是用双重筛选法，结果从 18 个原始特征中选出 5 个特征。

1. 数据处理步骤：

(1) 数据规格化：原始数据是样本中各个细胞的参数值，单位不同，数值大小相差很大，为便于处理先将原始数据规格化，公式是：

$$x_i' = \frac{x_i - M_k}{\Sigma_k}$$

其中：x_i'—规格化后特征数值，

x_i—规格化前特征数值，

M_k—各特征两类在一起的平均值，

Σ_k—各特征两类在一起的方差值

(2) 分别计算两类中各特征的平均值及协方差矩阵。

(3) 计算各特征的 Fisher 系数。

(4) 求两类投影到一维后的数据及平均值，这里只列出各组主要数据。

表 1 训练集样品中各类细胞主要数据(已规格化)

特征编号			1	2	3	4	5
平均值	4μ	正常	−0.741	−0.897	−0.144	−0.638	−0.185
		异常	0.419	0.506	0.081	0.360	0.105
	1μ	正常	−0.520	0.584	0.227	0.250	−0.335
		异常	0.142	−0.159	−0.062	−0.068	0.091
Fisher 系数	4μ		15.53	−12.99	2.11	−8.98	1.87
	1μ		2.09	4.52	−0.651	2.47	0.512

投影到一维后的平均值：

4μ：正常细胞 $D_1 = 5.216$

异常细胞 $D_2 = -2.945$

1μ：正常细胞 $D_1 = 1.854$

异常细胞 $D_2 = -0.51$

2. 判决阈值的选择及结果

如果不考虑两种错误的损失不同，可用两类投影后平均值的平均数作为判决阈值。例如：

4μ 的判决阈值为

$$D_0 = \frac{35 \times 5.216 - 62 \times 2.945}{35 + 62} = 0$$

1μ 的判决阈值为

$$D_0' = \frac{12 \times 1.854 - 44 \times 0.51}{12 + 44} = -0.88$$

按此阈值判决分类结果为：

步距 4μ 正常细胞误识率为 $\frac{2}{35} = 5.7\%$

异常细胞误识率为 $\frac{5}{62} = 8\%$

步距 1μ 正常细胞误识率为 $\frac{1}{12} = 8\%$

异常细胞误识率为 $\frac{9}{44} = 20\%$

可以看出与用双重筛选判决结果基本相同，对异常细胞的误识率(把异常细胞误分为正常细胞，即假阴性)都很大。但一般希望对异常细胞的误识率要小，而对正常细胞的误识率则可以大些，故希望能在给定假阴性率的条件下

表 2　步距 4μ 异常细胞

（平均值 −2.945　方差 3.05）

区　间	−8—−6	−6—−4	−4—−2	−2—0	0—2	2—4	4—6
细胞数	7	14	21	14	2	0	4

表 3　步距 4μ 正常细胞

（平均值 5.216　方差 2.65）

区　间	−5.5—−2.5	−2.5—0.5	0.5—3.5	3.5—6.5	6.5—9.5	9.5—12.5
细胞数	1	2	4	17	10	1

表 4　步距 1μ 异常细胞

（平均值 −0.51　方差 1.5）

区　间	−5—−4	−4—−3	−3—−2	−2—−1	−1—0	0—1	1—2	2—3	3—4
细胞数	1	1	2	13	17	4	2	2	3

表 5　步距 4μ 异常细胞（另一组特征）*

（平均值 −2.8　方差 2.5）

区　间	−7.8—−5.8	−5.8—−3.8	−3.8—−1.8	−1.8—0.2	0.2—2.2	2.2—4.2	4.2—6.2
细胞数	5	16	25	10	3	2	1

* 另一组特征指从原来 189 个特征中取另外 5 个特征，经 Fisher 法投影后所得数据，步距 1μ 的正常细胞只有 12 个，未画直方图。

选择判决阈值，为此对各类的分布规律做了进一步研究。

首先选定了 5 个特征后，粗略描绘了两类分别对各个特征的直方图，发现基本都是单峰的，但很多是不对称的，可见尚不能认为两类的分布是多维正态的。

用 Fisher 投影成一维分布后，进一步研究两类投影后的一维分布，先把各类数据按等间隔区间画出直方图，同时计算方差的估计值，结果如表 2—表 5，图 1—图 4，据各组的平均值及方差把理论计算的正态分布也画在同一图上（用点表示）作为比较。

从各图看，各分布比投影前更接近正态分布。

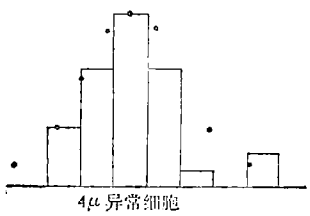

图 1　4μ 异常细胞

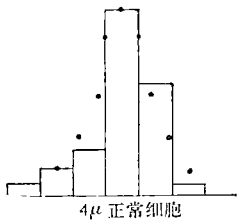

图 2　4μ 正常细胞

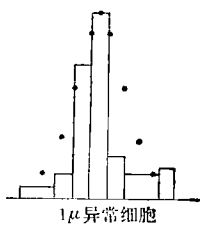

图 3　1μ 异常细胞

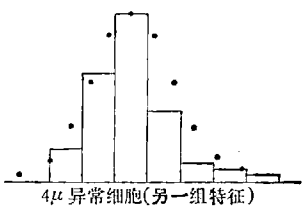

图 4　异常细胞（另一组特征）

进一步取表 1 数据用 x^2 检验来检查与正态曲线的适合程度[2]，为此先把横坐标分成等概率区间，统计各区间细胞数，同时把理论上应包含的细胞数也算出来列于表 6。

计算 x^2 统计量时把 1、2 二组 10、11 二组

表 6　4μ 异常细胞

（平均值 −2.945　方差 3.05）

区　间	理论上应包含的细胞数 p_i	实际出现的细胞数 f_i
$-\infty$——7.96	3.1	1
−7.96——6.1	6.2	6
−6.1——5.0	6.2	8
−5.0——4.12	6.2	7
−4.12——3.33	6.2	9
−3.33——2.56	6.2	7
−2.56——1.77	6.2	5
−1.77——0.82	6.2	6
−0.82—0.21	6.2	7
0.21—2.07	6.2	2
2.07—∞	3.1	3

都合并计算：

$$\chi^2=\sum\frac{(p_i-f_i)^2}{p_i}=\frac{(9.3-7)^2}{9.3}+\frac{(8-6.2)^2}{6.2}+\frac{(7-6.2)^2}{6.2}+\frac{(9-6.2)^2}{6.2}+\frac{(7-6.2)^2}{6.2}+\frac{(5-6.2)^2}{6.2}+\frac{(6-6.2)^2}{6.2}+\frac{(7-6.2)^2}{6.2}+\frac{(9.3-5)^2}{9.3}=4.8$$

数据共 9 组，因均值及方差都是从实验数据计算的估计值，故自由度为 9 − 3 = 6。查表得此组数据来自正态分布的置信水平约为 60%。说明投影后比较接近正态分布。细胞分类时一般要求假阴性率为 3—5%，而假阳性率允许达 20—30%，这时应按 Neyman-Pearson 规则去选判决阈值，假阴性率满足一定值条件下使假阳性率最小，但对两类且分布是正态时问题变得很简单，假阴性率给定后假阳性率也随之一定了。阈值计算过程如下：

步距为 4μ 时已知异常细胞分布的平均值 −2.945，方差 $\sigma=3.05$。假阴性率给定 3% 时，据正态分布表计算阈值 d 如下：

$d=\mu+1.88\sigma=-2.945+1.88\times3.05=2.43$，据此再计算理论上假阳性率的数值：正常细胞平均值 5.216，$\sigma=2.65$，判决阈值与正常细胞平均值的距离为：

$5.216-2.43=2.79$ 而 $\frac{2.79}{2.65}=1.05$

查正态曲线表得假阳性率应为 30%

按 $d=2.43$ 对 4μ 数据分类结果如下：

异常细胞误识率（假阴性率）$=\frac{4}{62}=6\%$

正常细胞误识率（假阳性率）$=\frac{4}{35}=11.4\%$

步距为 1μ 时阈值如下：

$d=\mu+1.88\sigma=1.88\times1.5-0.51=2.31$

按此阈值对 1μ 数据分类结果为：

异常细胞误识率 $=\frac{4}{44}=9\%$

正常细胞误识率 $=\frac{7}{12}=58\%$

由计算结果看出步距小了误识率反而加大。这是由于 4μ 的对象是区分正常与异常细胞，本较容易，1μ 的对象是区别异常细胞中的两种异常类型，原较困难。从数据上看，4μ 数据两类平均值的距离为 2.945 + 5.216 = 7.17，而 1μ 数据两类平均值的距离为 1.85 + 0.51 = 2.36，后者距离要近的多。

三、讨　论

1. 从上述结果看，虽然按给定假阴性率选择阈值，异常细胞的误识率仍较高。为此重新查对了误识的几个细胞，以 4μ 的数据为例，被误识的四个细胞投影后的平均值分别为：5.03（第 15 号细胞），4.25（121 号），4.71（138 号）和 5.72（175 号），均与正常细胞的平均值（5.216）靠的很近，反映在直方图上（图 1）也很明显，而与异常细胞平均值却离的较远。出现这一现象的原因是：即使同是正常或异常细胞，但因所处位置（浅层、底层）不同，其参数变化很大。这次样品中所选的正常细胞是中浅层的，异常细胞是底层的，因此中浅层的重增细胞（如 15 号）很容易判成正常细胞，而表层的退化细胞（175 号）也很接近正常细胞，而 121 号细胞经再次核对发现确系正常细胞，138 号细胞因与另一细胞靠的太近，数据在扫描时就有很大失真。以

上都说明对训练集的样品应严格检查，并应采取分层分类，以减小误识率。

2. 实验结果证明，用 Fisher 法把多维分布投影成一维分布后，更加接近正态，这时可按给定的假阴性率选择判决阈值，以降低对异常细胞的误识率。但投影后的分布与正态分布的适合程度如何，由于现有数据太少，尚难肯定。

参 考 文 献

[1] Duda, R. O. and Hart, P. E.: "Pattern Classification and Scene Analysis", Chap. 4, 1973.

[2] 中国科学院计算中心概率统计组：《概率统计计算》，科学出版社，1979年，第一章，第4节。

结　　语

以上介绍全部工作中，除第一个专题是为整个研究工作提供可以满足机器识别特殊要求的细胞涂片，其余四个专题提出了数字图象的预处理、特征抽提、特征选择、分类判别函数，再加上各个步骤间相互连接的软件，所有这些内容就构成了一个完整的算法系统。掌握了这一算法系统，标志着对模式识别概率统计方法全过程认识有了良好开端，并为开展下步工作打下了可靠的基础。有了这个基础，只要具备必要的硬设备，食管癌细胞的自动分类工作就立刻可以应用于临床；同时对其他种类细胞进行自动分类的应用性研究工作也就较易开展，因为完整的算法系统已为按照给定的指标（即模式识别术语的"特征"）进行细胞自动分析和类别判决提供了现成的贮存数字化图象的数据库和分析图象的程序包。从提高的角度来看，有了这一基础就可以进一步开展有关扩大判别、分类特征种类，以及分层判别、增加分类类别以提高识别率等研究工作，使这种新的分析方法可以解决生物学现有的常规方法所无法解决的生物医学图象分类问题。考虑到现已掌握的只是算法系统的最基本内容，这种提高工作就更加必要，而这正是我们今后工作的重点。

为了取得更多的成果，在研究中学科间相互渗透、紧密配合是十分重要的。回顾开展工作的初期，往往甲提出的问题乙不理解，乙阐述的概念甲又感到陌生。但是在共同的目标下，经过相互学习，共同实践，彼此之间的共同语言愈来愈多，工作进展也随之加快。如果今后从事这方面研究的工作者在开展工作时更自觉地注意取得共同语言，同时争取在国际上有一定水平的我国数学界的有力配合，以发展新概念，提出新方法，并密切结合我国所特有的研究对象，虽然我国起步较晚，硬设备也有一定限制，但可以预期将作出具有我国特色的工作来。

[本栏各文于 1979 年 12 月 12 日收到]

学 科 简 介

细 胞 图 象 分 析

模式识别（Pattern Recognition）是近二十年来随着电子计算机的广泛应用而产生和发展的新兴学科。它企图用计算机连同特殊输入装置模拟或部分代替人类的视觉、听觉或其他智能，以分析、描述与识别图象、声音或其他模式。这一学科当前在机器人视觉，生物医学图象处理、遥感图象的解释和文字、声音识别等方面均有广泛应用并取得了卓越的成果。

应用模式识别方法于细胞学研究实际上是将现代技术中的信息理论用于分析细胞。它既可科学地总结临床细胞病理学家的诊断经验，又可充分发挥机器视觉分辨率高（可达 256 灰阶）。抽提特征方式灵活多样（可按各种数学模型形成特征，不一定要求每个特征具有明确的直观意义。例如美国对一细胞可提取 200 多个特征，加拿大可提取 300 个以上）。随着细胞图象自动分析方法的完善与工业基础的发展，不难实现癌细胞的自动筛选与诊断，同时也可为疑难病症的鉴别诊断和对从癌前到早期癌演变过程的研究提供客观指标与有效手段。

癌细胞学自动化与细胞图象分析第三次国际会议今年五月将在西德慕尼黑召开。

（中国科学院生物物理研究所陈传涓）