逐步判别分析在疟原虫血涂片细胞分类中的应用

丁 岩 柴振明
(中国科技大学研究生院电子学部,北京) (中国科学院电子学研究所,北京)
陈传涓
(中国科学院生物物理所,北京)

提 要

本文应用逐步判别分析方法对疟原虫血涂片细胞进行了分类研究。用169个红细胞(其中正常红细胞为130个,带疟原虫的红细胞为39个)作为训练集,192个红细胞(其中正常红细胞为157个,异常的为35个)作为考试集进行了统计分析,并通过实验调整判别阈值。对考试集的判别:假阴性率为11.4%,假阳性率为7.6%,结果较为理想。

关键词:细胞图象处理,特征抽取,逐步判别,模式识别

一、前 言

疟原虫血涂片细胞的计算机自动分类是疟疾普查工作中提出的需迫切解决的课题.目前,疟疾的普查方法是由医务人员显微镜下观察血涂片细胞图象进行诊断,此项工作需要有丰富实践经验的医务人员,而且工作量很大。计算机和模式识别技术为我们提供了用计算机自动分析血涂片以代替人工观测的可能性。疟疾是第三世界国家的流行病[1],西方国家对疟原虫血涂片细胞的自动分析判读研究较少。因此,"疟原虫血涂片细胞的计算机自动分类"研究具有重要的实际意义。

二、系统设计

我们的工作是在通用图象分析仪Magiscan 2上进行的。分析的疟原虫血涂片采用姬氏染色,图1(见封三)是经显微镜放大后的细胞图象,图2(见封三)为输入计算机后又经软件放大后的细胞图象。其中带有斑点和环状体的为寄生有疟原虫的红细胞,其余为正常红细胞。

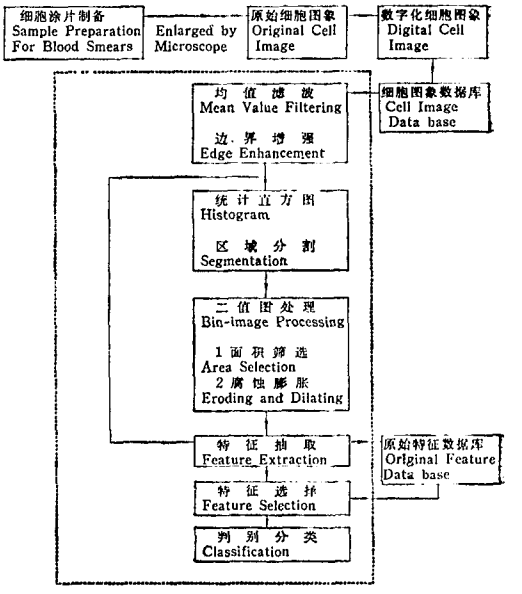

图3 疟原虫血涂片细胞图象识别系统方框图
Fig.3 A System block diagram for the classification of plasmodium blood smear

对疟原虫血涂片细胞图象进行处理和识别

的结构框图如图 3 所示[2]，虚线内为程序库。

(1) 细胞涂片制备：姬氏染色，要求涂片内的细胞尽量散开，避免重叠；其次要求背景清晰干净，染色要标准。

(2) 细胞涂片经显微镜放大(100×20)后，得到原始细胞图象；又经 A/D 转换数字化后，得到 512 × 512 点阵 2^6 灰值数字图象，从而建立细胞图象数据库。

(3) 对细胞图象数据库中的图象进行均值滤波，边界增强等预处理。

(4) 对灰值图象统计直方图并查寻谷点，用阈值法进行区域分割，将细胞从背景中分离出来，得到二值图象。

(5) 对二值图进行处理。面积筛选以剔除面积较小的杂质和面积较大的白细胞；腐蚀膨胀以清除噪声点，并锐化细胞边缘。

(6) 特征抽取：选择最能表征正常红细胞和异常红细胞差异性的特征，并建立原始特征数据库。

(7) 选择最有效的特征并进行判别分类。

参考医生的经验并经过初步分析，我们抽取了下面九个细胞灰值和几何特征作为逐步判别分类算法的特征矢量[3]。

(1) x_1：细胞平均光密度；

(2) x_2：细胞光密度方差；

(3) x_3：细胞光密度变化系数；

(4) x_4：细胞积分光密度；

(5) x_5：细胞最大光密度与最小光密度之比；

(6) x_6：细胞面积；

(7) x_7：细胞周长；

(8) x_8：细胞形状因子；

(9) x_9：细胞轮廓中的凹点数。

三、逐步判别法基本思想

逐步判别的基本思想，即每一步选一个判别能力最显著的自变量进入判别函数。而且在每次选变量之前都对已经进入判别函数的诸变量逐个检验其显著性，如果发现有某个变量由于新变量的引入而变得不“重要”(即它在判别函数中判别能力不显著)时；就剔除这个变量，直到判别函数中包含的所有变量判别能力都显著为止。因此，这样得到的判别函数，它所包含的变量都是“重要”的[4]。

逐步判别不仅可以节省大量的计算工作量，从众多的自变量中选出最重要的变量构成判别函数，而且可以增加判别函数的稳定性，从而提高判别的效果。逐步判别法的原理是基于正态母体等协方差的样本分布而推导的，我们通过直方图分析了每个特征的分布，发现大多数样本是单峰对称的，少部分是单峰不对称的。因此可假定其为正态分布，并进行了如下实验。

用 169 个红细胞作为训练集（其中 130 个为正常红细胞，39 个为寄生有疟原虫的红细胞)，设 $q_g(g=1,2)$ 为细胞的先验概率，通过计算[5]得到判别系数 C_{0g}，$C_{ig}(i=1,2,\cdots,9;\ g=1,2)$。从而得到判别函数：

$$y_g(\boldsymbol{x}) = \ln q_g + C_{0g} + \sum_{i=1}^{9} C_{ig}x_i$$

若 $y_g^*(\boldsymbol{x}) = \max\limits_{1\leqslant g\leqslant a}\{y_g(\boldsymbol{x})\}$ 则把 $\boldsymbol{x}$ 划归为第 g^* 个母体。这里 a 为分类母体数目，在此 $a=2$。

四、计算结果

设定如下参数：

a：分类数；　M：自变量个数；

N：总观测次数；

$F1$：用作选入的 F_a 统计量；

$F2$：用作剔除的 F_a 统计量。

输入以下参量，$a=2$，$M=9$，$N=169$，$F1=2.0$，$F_2=2.0$，进行计算，得到如下结果(表 1 所示)。

其中，L 为已引入变量，u 为最终的 Wilks 量，x^2 为 u 的近似检验。查表得 $x^2_{0.005}(4)=14.86$ $x^2=1.62792\times10^2>x^2_{0.005}(4)$。故选出的 4 个特征对两个母体有非常显著的区分能力。

所取特征量分别为：

x_5：细胞最大光密度与最小光密度之比；

x_8：细胞形状因子；

表1 Table 1

Step	1	2	3	4
Variables of selected or replaced	$+x_5$	$+x_8$	$+x_4$	$+x_7$
F test	1.00294×10^{2}	7.32987×10^{1}	1.98334×10^{1}	6.18695
Selected variables	$1(x_5)$	$2(x_5,x_8)$	$3(x_5,x_8,x_3)$	$4(x_5,x_8,x_3,x_4)$
u	6.24780×10^{-1}	4.33406×10^{-1}	3.86900×10^{-1}	3.72835×10^{-1}
x^2	7.83142×10^{1}	1.38789×10^{2}	1.57157×10^{2}	1.62792×10^{2}

表2 Table 2

i \ j	0	4	5	7	8
1	−10.1732	−6.00206	26.8894	4.56840	−0.78162
2	−10.0668	−9.07343	−5.04787	6.06671	−0.33776

x_4: 细胞积分光密度;

x_7: 细胞周长。

分别表征了细胞的灰值和几何信息。

我们假设两类样本的先验概率分别为其在训练集中所出现的频率，根据 Bayes 准则可得线性判别函数。

最后运算产生的判别系数 $c[i, j]$ 为表 2 所示。

除了表定义的 $c[i,j]$ 外，其余 $c[i,j]=0$。

由上面判别系数对训练集 169 个样本进行判别得到表 3 所示结果。

表3 Table 3

Computational classification \ Original classification	1	2	Add together
1	32	6	38
2	7	124	131
Add together	39	130	169

由此可计算出假阳性率为: 6/130=4.6%，假阴性率为: 7/39=17.9%。可见假阴性率较高，假阳性率较低，而一般观测血涂片的要求是限制假阴性率在一定范围内，为此我们可以调整判别函数值，以改变判别效果。

由前面可知判别函数为: $y_g(\boldsymbol{x})=\ln q_g+c_{0g}+c_{1g}x_1+\cdots+c_{9g}x_9$，我们可以通过调整 c_{0g} 来改变判别结果。对寄生有疟原虫的红细胞，$c[1,0]=-10.1732$，对正常红细胞，

表4 Table 4

Result \ $c[1,0]$	1)	False positive rate	2)	False negative rate	3)	Error rate
−5.1732	31	23.0%	1	2.6%	32	18.9%
−6.1732	14	10.8%	2	5.1%	16	9.5%
−7.1732	8	6.2%	3	7.7%	11	6.5%
−8.1732	6	4.6%	5	12.8%	11	6.5%
−9.1732	5	3.0%	8	20.5%	13	7.7%
−10.1732	3	2.3%	11	28.2%	14	8.3%

1): Numbers of normal red cells which are recognized as abnormal ones

2): Numbers of abnormal red cells which are recognized as normal ones

3): Numbers of red cells which are false recognized

表5 Table 5

c[1,0] \ Result	1)	False positive rate	2)	False negative rate	3)	Error rate
−5.1732	49	31.2%	0	0.0%	49	25.5%
−6.1732	22	14.0%	2	5.7%	24	12.5%
−7.1732	12	7.6%	4	11.4%	16	8.3%
−8.1732	6	3.8%	5	14.2%	11	5.7%
−9.1732	4	2.5%	7	20.0%	11	5.7%
−10.1732	3	1.9%	9	25.7%	12	6.3%

1),2),3): The same meaning as the table 4

$c[2,0] = -10.0668$。表 4 是改变 $c[1,0]$后所得到的测试结果，表中正常红细胞个数为 130，异常的为 39，总细胞个数为 169。

由表 4 可见，当$c[1,0]$为−7.1732 时，假阴性率和假阳性率都达到相对较小的值，分别为 7.7% 和 6.2%。表 5 为改变$c[1,0]$时，对考试集 192 个样本测试所得到结果(其中正常红细胞个数为 157，异常红细胞数为 35)。

由表 5 可以看出，对考试集的判别，当$c[1,0]$为−7.1732 时，假阳性率和假阴性率也相对较低，分别为 7.6% 和 11.4%。因此，在判别系数 $c[i,j]$ 中，令 $c[1,0]$ 为−7.1732，其它保持不变，能得到更为满意的判别效果。

五、讨　论

从上面分析可以看出，用逐步判别法对疟原虫血涂片细胞进行分类，能得到较为满意的结果。因此，假定特征矢量近似服从正态分布也是成立的。用于判别的四个特征，x_5：细胞最大光密度与最小光密度之比，x_8：细胞形状因子，x_4：细胞积分光密度，x_7：细胞周长，分别表征了细胞的灰度和几何信息。根据有经验的医务人员的判断，被寄生红细胞的变化为：细胞胀大，色淡，常呈长圆形或多边形，滋养体期开始出现鲜红色的薛氏点。可见，用逐步判别法选择的特征与医生的经验基本相符。我们还另抽取了细胞纹理特征，以期提高正确识别率，但效果不明显，而且需更长时间的预处理和特征计算。

寄生于红血球中的疟原虫，其发育过程大致经历五个时期，即环状体、大滋养体、裂殖体前期、成熟裂殖体和配子体时期，各时期的形态差异较大。对错判细胞的分析表明，寄生有疟原虫的红细胞错判为正常红细胞，主要是由于寄生的疟原虫尚处于环状体期，形态较小，染色后，着色点也较小。正常细胞判为异常的，主要是细胞局部染色不匀和"噪扰"所造成的。若能使涂片染色更为均匀，并减少杂质，抽取细胞的颜色信息，可望进一步提高正确识别率。

国际上，细胞的计算机自动分类工作已有三十多年的历史，并且已有用于白血球分类和癌细胞识别的商用机投入市场[6,7]。通过技术人员的目识和机器自动识别的比较表明，技术人员之间计数的差别等于或大于机器计数和技术人员的平均结果之间的差别。这说明不能达到精确一致的主要因素是统计"噪扰"，而不是机器误差，这些商用机器已确实能够代替技术人员的工作。

上述工作在我国尚处于初级阶段，以前虽曾有人对白细胞和癌细胞的自动识别作了一些工作，但大都限于软件算法系统的研制。本文工作也是在软件算法系统上作了一些尝试，若要真正达到实用，还有待进一步工作，如硬件机器及相应设备，自动涂抹机和染色设备、皮带传送系统的研制。

衷心感谢北京医科大学寄生虫病教研组和中国医科院基础研究所寄生虫病研究室对本文工作所给予的支持。感谢中科院电子所的高宇同志对本文工作所给予的帮助。

参 考 文 献

1 中山医学院. 人体寄生虫学; 北京: 人民卫生出版社, 1979; 76--94
2 丁岩,柴振明,陈传涓. 北京生物医学工程,1988;7(2): 50
3 Ashoky Kulkarni. *The Journal of Histochemistry and Cytochemistry*, 1979;27(1): 28,39
4 南开大学数学系统计预报组.概率与统计预报及在地震与气象中的应用. 北京: 科学出版社,1978;157--177
5 中科院计算中心概率统计组. 概率统计计算. 北京: 科学出版社,1979;192—206
6 Norgren P E. *Pattern Recognition*, 1981; **13** (4): 299,314
7 Fu K S. *Applications of pattern recognition*, Boca Raton, CRC Press, 1982: 184—194

[本文于 1989 年 1 月 3 日收到]

THE APPLICATION OF STEPWISE DISCRIMINANT ANALYSIS IN THE CLASSIFICATION OF PLASMODIUM BLOOD SMEAR

Ding Yan
(Graduate School, Academia Sinica, Beijing)

Chai Zhenming
(Institute of Electronics, Academia Sinica, Beijing)

Chen Chuanjuan
(Institute of Biophsics, Academia Sinica, Beijing)

ABSTRACT

An algorithm of Stepwise Discriminant Analysis is used in the classification of Plasmodium blood smear. 9 features of the red cell is extracted. 169 red cells (among them 130 cells are normal, 39 contain Plasmodium) are used as training set, 192 red cells (among them 157 cells are normal, 35 contain Plasmodium) as test set. We have done the statistical analyses and got good results. For test set, the false negative rate is 11.4%, the false positive rate is 7.6%.

Key words cell image processing, features extraction, stepwise discriminantion, pattern recognation

---

(上接第120页)

元素与人类健康讨论会"总结来看,苏联对今后微量元素的发展似可归纳为: (1)强调宏观与微观相结合,各种层次(整体,器官,组织,细胞,亚细胞,分子)相结合。会议对我国已从线粒体(亚细胞)水平来研究克山病的发病机理给予高度的评价。(2) 加强微量元素的基础研究, (3) 鼓励与支持新技术新方法的探索与应用。

我国近年来对微量元素研究开展了不少工作, 取得了很多成绩,尤其是微量元素硒,无论在结合大骨节病、克山病、肝癌的发病机理与预防, 还是在基础研究方面所取得的成绩都引起国际上很大的关注。"全苏微量元素与人类健康讨论会" 所提出的一些意见与问题对我们也有一定的参考价值。今后,我国微量元素应继续发扬在硒研究方面的优势,使之不断系统与深入,以便获得更高水平的成果。与此同时, 除硒以外的其它微量元素的研究也应逐步开展与加强。对微量元素与人类健康的研究,注意将不足,过量与平衡失调的问题予以全面考虑。此外,微量元素的应用研究(微量元素与农业、医药以及工业污染)和基础研究应有一个合理的安排。为了加强这方面的领导与协调, 除成立相应的学会 (中国营养学会下设的微量元素学会即将成立)外, 建议卫生部、中国医学科学院或中国预防医学科学院成立微量元素委员会。

[中国科学院生物物理研究所 杨福愉]

丁　岩等："逐步判别分析在疟原虫血涂片细胞分类中的应用"一文的附图1，2

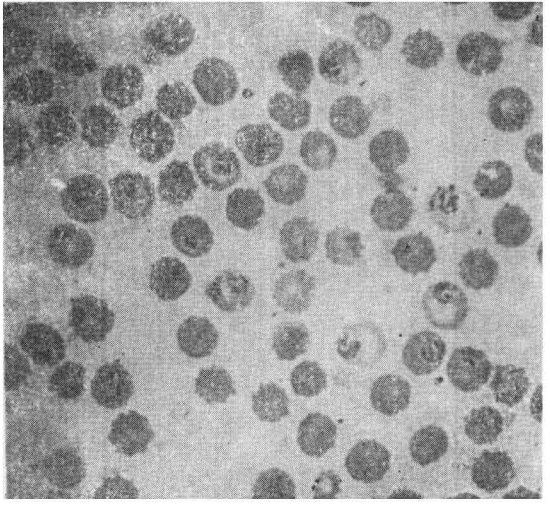

图1　经显微镜放大后的细胞图象

Fig. 1　Cell image enlarged by microscope

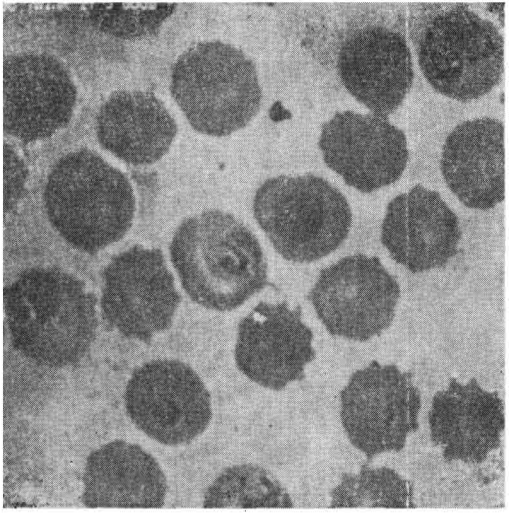

图2　又经计算机软件放大后的细胞图象

Fig. 2　Cell image enlarged again by computer software