第12卷 第1期 自 动 化 学 报 Vol. 12, No. 1
1986 年 1 月 ACTA AUTOMATICA SINICA Jan., 1986

类间散度特征选择法及其应用

钱 学 双
(中国科学院自动化研究所)

摘 要

本文介绍了统计模式识别中特征选择的一种方法——类间散度法. 它是以两类细胞样本在 n 维特征空间中 Fisher 映射点集的类间差与类内差之比作为特征选择的判据. 用这种方法对五类白血球进行了特征选择，并对所选特征用训练集和测试集进行了识别效果检验，同时，得到了特征选择的判据 D 与识别率之间的关系曲线.

一、引 言

特征选择是图象识别中的一个关键问题[1]. 为了全面描述一个图象，要尽可能多地抽取特征;为了识别图象,还希望用最少的特征建立与其他事物相区别的判别模式，以得到最好的分类结果. 一个比较好的分类器的建立,除了本身需具有最佳设计方案外,还依赖于所选用的特征子集. 本文介绍的特征选择方法以类间散度作为选元的判据. 类间散度是两类母体样本在 n 维特征空间中 Fisher 映射点集的类间差与类内差之比[2]. 比较各特征子集的类间散度的大小以决定特征参量的取舍.

在白血球的计算机自动分类研究工作中，已将这种方法用于五类细胞的特征选择和分类判别. 根据文献[3]描述的白血球的 187 个特征，对白血球中的嗜中性(NEU),淋巴(LYM)，单核 (MON)，嗜酸 (EOS)，嗜硷 (BAS) 五类细胞共 1330 个子样,按照图 1 树状分类器进行了特征选择,同时,用一半子样作为训练集,另一半作为测试集,进行了识别效果检验,得到了反映整个识别状况的两个混淆矩阵，取得了识别率分别为 89.9% (训

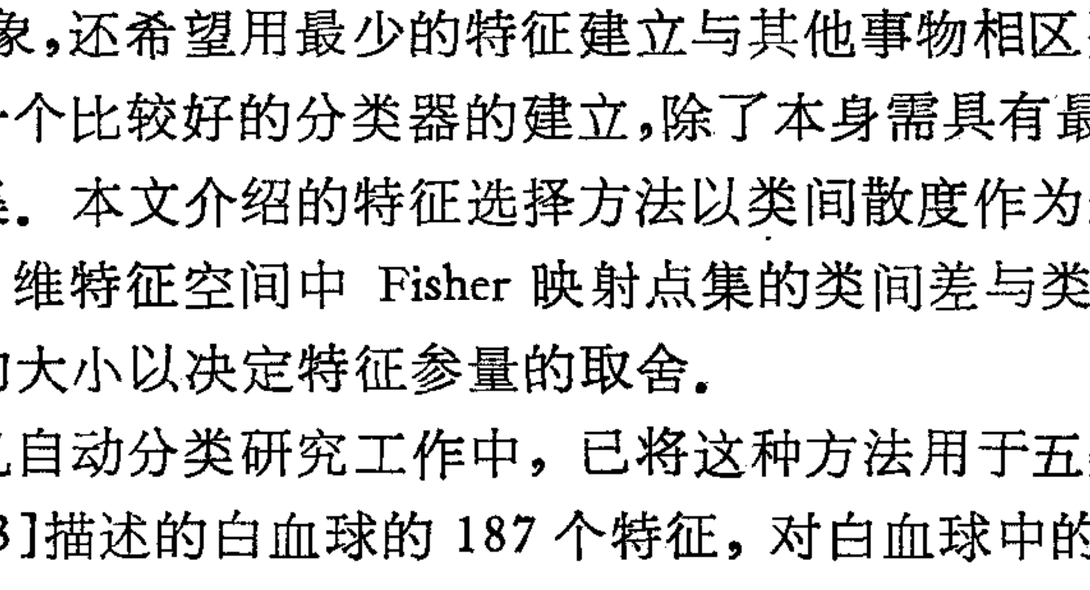

图 1 五类细胞树状分类器

本文于 1984 年 2 月 22 日收到.

练集)和 86.8%(测试集)的分类结果,并且找到了特征选择判据 D 与识别率之间的非线性关系曲线.

二、特征选择判据、步骤及程序框图

1. 特征选择的判据——类间散度 D

设 P, Q 两类细胞子样分别为 x_{lj}^p, x_{tj}^q ($l=1,\cdots,m_p$; $t=1,\cdots,m_q$; $j=1,\cdots,n$). m_p 和 m_q 分别为 P, Q 两类细胞的样本数, n 为特征个数. 根据 Fisher 映射,两类细胞样本在 n 维特征空间中向一维空间投影,其映射向量[4]为

$$\boldsymbol{V}=\frac{1}{2}(\Sigma^p+\Sigma^q)^{-1}(\bar{X}^p-\bar{X}^q). \tag{1}$$

这里, Σ^p 和 Σ^q 分别为 P, Q 两类细胞样本的离差矩阵; $\bar{X}^p$ 和 $\bar{X}^q$ 为各自的均值向量. 两类细胞母体样本的 Fisher 映射函数分别为

$$y_l^p=\sum_{j=1}^{n}V_jx_{lj}^p,\ (l=1,\cdots,m_p) \tag{2}$$

$$y_t^q=\sum_{j=1}^{n}V_jx_{tj}^q.\ (t=1,\cdots,m_q) \tag{3}$$

式(2),(3)构成了两类细胞母体样本的映射点集.

设 M_p 和 M_q 为两类细胞样本映射点集的均值, S_p 和 S_q 为映射点集的标准差,则有

$$D=|M_p-M_q|/(S_p+S_q). \tag{4}$$

式(4)右边的分子部分为映射点集的类间差,分母部分为类内差. 显然, 对不同的特征集, D 愈大,分类效果应愈好. D 反映了两个模式间的距离,在多维空间中,则反映两类或多类模式的分散度. 为了与分类概念联系起来,将 D 称作类间散度,或简称散度.

2. 特征选择的过程

根据不同的特征组合所得到的散度, 选择类间散度最大的那个特征组作为最佳特征子集,以此建立区分两类细胞图象的判决方程,具体步骤如下:

在进行选择之前,可以粗略估计一下最多可选用的特征个数. 对于一个判决方程,特征参量一般不要超过十个. 最终选定的特征的多少, 可通过比较类间散度 D 的大小来确定.

首先,用穷举法选出最初的两个特征. 每一次从 N 个特征中, 抽出一个特征 j, 找出两类细胞样本在该特征下各自的均值 $\bar{x}_j^p$, $\bar{x}_j^q$ 和标准差 s_j^p, s_j^q, 并通过公式

$$d_j=|\bar{x}_j^p-\bar{x}_j^q|/(s_j^p+s_j^q) \tag{5}$$

得到这个特征的类间散度,由此, 可以求得 n 个单一特征下的类间散度 d, 找出其中最大的两个 d_{k_1}, d_{k_2}, 它们所对应的第 k_1, k_2 两个特征 f_{k_1}, f_{k_2} 即是最初选入的特征. 当然,随着后来新的特征的引入,这最初入选的两个特征可能继续留用, 也可能被剔除. 当前, 这两个特征暂时构成了优选子集,记为 F_k.

第二步,从不包括第 k_1, k_2 两个特征的集合中(含 $n-2$ 个特征),分别抽出一个特征 $f_J*[f_J*\notin(f_{k_1},f_{k_2})]$, 与最初入选的两个特征 f_{k_1}, f_{k_2} 组成新的特征集. 根据公式(1)到(4), 找出新特征集下的类间散度. 从所得到($n-2$ 个)类间散度中, 找出最大的记为

D^*，它所对应的特征记为 f^*.

第三步，引入了新的特征参量之后，还要确定最初入选的两个特征在新的特征子集中的地位. 从三个特征中任意删掉一个，用其余两特征按上面的步骤进行映射变换，从而可得到 C_3^2 个两特征下的类间散度，找出其中最大的，记为 D'，此时删掉的那个特征记为 f'. 如果 f' 等于 f^*，这说明最初入选的两个特征还可以继续留用，由新引入的特征 f^* 与之构成的特征组合当前分类效果最好，记 $f_{k_3}=f^*$. 此时最佳特征子集为 $F_k\{f_{k_1}, f_{k_2}, f_{k_3}\}$. 返回第二步，继续寻找下一个特征.

若 f' 不等于 f^*，而等于 f_{k_1} 或 f_{k_2}，则说明最初在单一特征下选入的两个特征之一应被淘汰. 此时，用 f^* 代替 f_{k_1} 或 f_{k_2}，形成新的两个特征组成的子集. 回到上一步，重新寻找新的 f^* 和 f'，直到 f' 等于 f^* 为止.

选入 $i+1$ 个特征后，若 $D_{i+1}<D_i$，则停止选元；若 $D_{i+1}>D_i$，则根据 ε 的值决定选元是否继续进行.

$$\varepsilon=(D_{i+1}-D_i)/D_i. \tag{6}$$

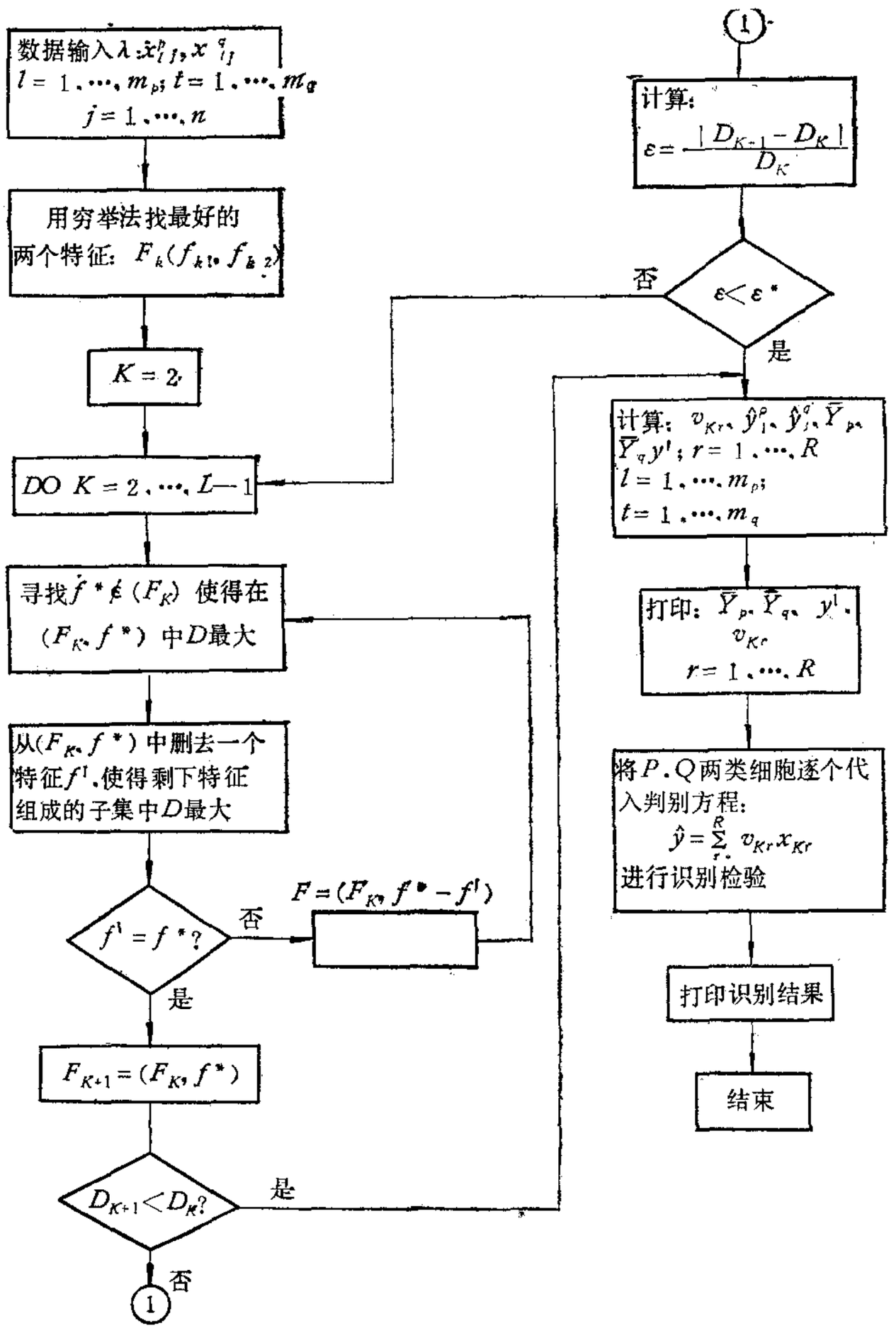

图 2 特征选择程序框图

若 $\varepsilon > \varepsilon^*$，则选元继续进行；若 $\varepsilon < \varepsilon^*$，则选元结束。$\varepsilon^*$ 根据选元情况设定，一般取 ε^* 为 0.01 到 0.05。用这种方法确定选元过程是否停止的理由是，若新增加一个特征，它对原有特征子集下的散度增加量不大，甚至小于前者，说明它的引入不能明显提高分类效果，故不予引入，选元到此结束。

若利用上面的方法已选出 R 个特征，并得到 R 个特征下的映射方程：

$$y_l^p = \sum_r^R v_{k_r} x_{lk_r}^p, \quad (l = 1, \cdots, m_p) \tag{7}$$

$$y_t^q = \sum_r^R v_{k_r} x_{tk_r}^p, \quad (t = 1, \cdots, m_q) \tag{8}$$

令 $\overline{Y}_p$，$\overline{Y}_q$ 分别代表 y_l^p 和 y_t^q 的均值，则有

$$y' = (\overline{Y}_p + \overline{Y}_q)/2. \tag{9}$$

y^1 即两类细胞分类的判决阈值。图 2 是实现以上步骤的程序框图，程序用 FORTRAN 语言编写。

三、特征选择计算结果

文献[3]描述的白血球的特征包括光密度参数、几何参数、纹理参数共 187 个。因计算机内存容量有限，不能将全部特征同时作为特征选择的基元，所以按特征的物理意义和数学含义，将它们大致分为六个特征组：1)光密度参数组 (24 个)；2)基本几何参数组(22 个)；3)傅氏系数组 (80 个)；4)灰度共生矩阵组 (20 个)，5)灰度游程长度矩阵组 (20 个)；6)一阶统计和十六个灰度梯度共生矩阵组 (21 个)。选取特征时，先从各个特征组选出一定数量的特征，然后再从这些入选的特征集中进一步选元。

全部细胞子样共 1330 个，其中嗜中性细胞 526 个，淋巴细胞 282 个，单核细胞170个，嗜酸细胞 254 个，嗜硷细胞 98 个。将各类细胞按自然序号排队，奇数序号的细胞作为训练集，偶数序号的细胞作为测试集。训练集和测试集的细胞子样各为 665 个。

为进行五类细胞的总分类，特征选择过程是在图 1 树状分类器中的各点上进行的。这种分类器的树状结构，是根据血液学专家和化验员对白血球的分类经验，并在参考其他几种选元方法的基础上产生的。因本文主要介绍的是特征选择方法，对分类器设计及最佳性问题这里不予讨论。

对每一节点上的两类细胞，按上面介绍的特征组分组选元，择优提取。经多次试验，第三组即傅氏系数组中的特征对分类效果作用不大，以后选元均不采用此组参量。

下面列出了图 1 分类器第 1，2 节点总的选元详细情况和第 3，4 节点总选元的基本情况。所列特征是在第 1，2，4，5，6 五个特征组中入选特征的基础上选元得到的。各特征小组选取特征情况见表 1。

第一节点：对嗜中性与淋巴、单核、嗜酸、嗜碱细胞进行分类。

入选特征(共八个)有：

10，14，27，132，137，169，180，182

映射方程为 $y = 29.72 \times 10^{-4} x_{10} + 16.99 \times 10^{-5} x_{14} - 19.58 \times 10^{-4} x_{27} + 20.32 \times$

表 1 树状分类器各节点不同特征组的选元情况

特征集 \ 类别	第一节点 P 类: LYM,MON EOS, BAS Q 类: NEU	第二节点 P 类: LYM,MON Q 类: EOS, BAS	第三节点 P 类: LYM Q 类: MON	第四节点 P 类: EOS Q 类: BAS
第一特征组 (1—24)	入选: 6, 8, 10, 14, 24 $D = 1.52$ 训练集: 93.1% 测试集: 91.8%	入选: 3,8,13, 16,18 $D = 0.86$ 训练集: 81.8% 测试集: 76.9%	入选: 27,29,33 37,39 $D = 1.23$ 训练集: 89.9% 测试集: 86.3%	入选: 26,31,32, 37,38 $D = 0.93$ 训练集: 81.3% 测试集: 79.5%
第二特征组(25—46)	入选: 27, 32, 34, 37, 38 $D = 1.17$ 训练集: 90.2% 测试集; 88.9%	入选: 26,27,36 38,45 $D = 0.99$ 训练集: 83.1% 测试集: 83.1%	入选: 27,29, 33,39,37 $D = 1.23$ 训练集: 88.9% 测试集: 86.3%	入选: 26,31, 32,37,38 $D = 0.93$ 训练集: 81.3% 测试集: 79.5%
第四特征组(126—145)	入选: 126,132, 134,137,138 $D = 1.68$ 训练集: 95.8% 测试集: 94.4%	入选: 130,135, 139,141,142 $D = 1.33$ 训练集: 90.5% 测试集: 87.6%	入选: 126,129, 130,135,143 $D = 1.14$ 训练集: 86.2% 测试集: 84.1%	入选: 127,133, 136,139,143 $D = 0.96$ 训练集: 82.9% 测试集: 81.2%
第五特征组(146—165)	入选: 149,151, 161,163,164 $D = 0.57$ 训练集: 73.1% 测试集: 71.7%	入选: 149,151 152,156,161 $D = 0.89$ 训练集: 82.8% 测试集: 79.6%	入选: 147,151 152,157,164 $D = 1.15$ 训练集: 86.4% 测试集: 85.4%	入选: 146,156 161,162,163 $D = 0.74$ 训练集: 75.6% 测试集: 65.9%
第六特征组(166—187)	入选: 169,179, 180,182,185 $D = 1.61$ 训练集: 94.0% 测试集: 92.9%	入选: 166,167, 177,182,187 $D = 1.04$ 训练集: 82.3% 测试集: 81.3%	入选: 169,175, 182,185,187 $D = 130$ 训练集: 90.3% 测试集: 88.5%	入选: 167,175, 177,185,186 $D = 1.11$ 训练集: 90.3% 测试集: 85.2%

$10^{-5}x_{132} - 30.51 \times 10^{-5}x_{137} + 26.21 \times 10^{-4}x_{169} + 10.42 \times 10^{-4}x_{180} - 29.57 \times 10^{-5}x_{182}$; 类间散度 $D = 1.882$; 阈值 $EY = 35.35 \times 10^{-3}$; 识别率: 训练集为 97.3%, 测试集为 96.7%.

第 2 节点: 对淋巴、单核与嗜酸、嗜碱细胞进行分类. 入选特征(共九个)有

8, 26, 38, 130, 135, 141, 142, 149, 177

映射方程为 $y = -81.51 \times 10^{-3}x_8 + 49.39 \times 10^{-4}x_{26} - 49.03 \times 10^{-4}x_{38} + 48.15 \times 10^{-3}x_{130} - 19.54 \times 10^{-3}x_{135} - 34.00 \times 10^{-2}x_{141} - 35.19 \times 10^{-5}x_{142} + 14.79 \times 10^{-6}x_{149} + 92.53 \times 10^{-4}x_{177}$; 类间散度 $D = 1.448$; 阈值 $EY = 31.26 \times 10^{-4}$; 识别率: 训练集为 93.0%; 测试集为 88.8%.

由于篇幅所限,第 3, 4 节点只介绍选元的基本情况.

第 3 节点: 对淋巴与单核细胞进行分类. 共选入十个特征: 8, 15, 19, 27, 37, 39, 130, 135, 169, 187. 类间散度 $D = 1.512$; 阈值 $EY = -13.38 \times 10^{-2}$.

第 4 节点：对嗜酸与嗜碱细胞进行分类，共选入八个特征：1，26，32，37，38，161，167，175．类间散度 $D = 1.530$；阈值 $EY = 43.18 \times 10^{-3}$．

四、总分类效果及类间散度与识别率的关系曲线

为了检验五类细胞总的识别效果，根据图 1 树状分类器及选元过程中确定的各节点上的判别方程、判决阈值，将训练集和测试集中的细胞(各有 665 个)，逐个代入进行识别率统计，得到反映训练集和测试集识别状况的两个混淆矩阵 (见表 2，表 3)．矩阵中对角线上的数为正确识别的细胞数，元素 $a_{ij}(i \neq j)$ 为第 i 类细胞错分到第 i 类中去的细胞数．

表 2　训练集混淆矩阵

机判 \ 人判	NEU	LYM	MON	EOS	BAS
NEU	256	4	1	5	1
LYM	1	120	2	4	2
MON	4	13	71	3	4
EOS	0	3	8	112	3
BAS	2	1	3	3	39

表 3　测试集混淆矩阵

机判 \ 人判	NEU	LYM	MON	EOS	BAS
NEU	257	6	1	6	3
LYM	1	118	10	11	4
MON	2	9	67	7	5
EOS	1	4	7	99	1
BAS	2	4	0	4	36

经过识别统计，训练集中分对的细胞数 (表 2 混淆矩阵中对角线元素的总和) 为 598 个，识别率为 89.9%．测试集中分对的细胞数(表 3 混淆矩阵中对角线元素的总和) 为 577 个，识别率为 86.8%．

最后，简要讨论特征选择判据 D 与识别率之间的关系．根据类间散度 D 的定义，在选取特征子集时，要求两类细胞样本类间差尽可能大，类内聚集性尽可能好，这是区分两类或多类模式的准则之一．为弄清特征选择判据 D 与识别率之间的关系，以类间散度为纵坐标，以识别率为横坐标，找出嗜酸与嗜碱细胞在不同散度下的识别率以及在坐标系中对应的点 $C_1(0.74, 75.6\%)$，$C_2(0.93, 81.3\%)$，$C_3(0.96, 82.6\%)$，$C_4(1.01, 83.5\%)$，$C_5(1.11, 89.0\%)$，$C_6(1.36, 92.6\%)$，$C_7(1.53, 94.3\%)$，用曲线逼近这些点，由此得到嗜酸细胞与嗜碱细胞的散度识别率关系图(图 3)．从图 3 中可以看到，识别率随类间散度增大非线性地升高，但在某一区间(D 在 1.1—1.5 之间变化时) 两者的关系近似于线性．

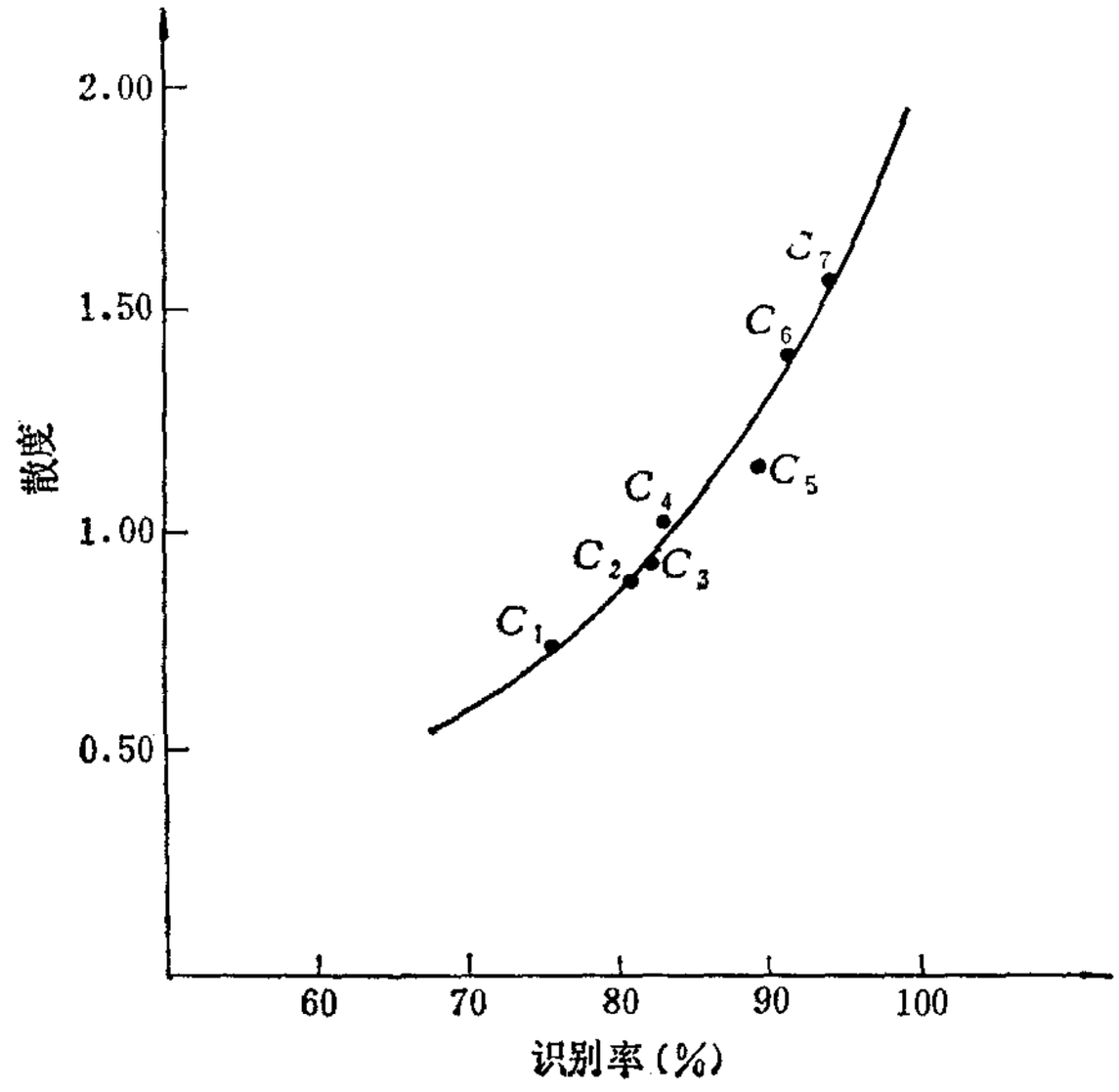

图 3 嗜酸、嗜碱细胞散度与识别率关系图

图 4 是分类器各节点上两类细胞不同特征组下的类间散度与识别率(在训练集内)的关系图. 由图 4 可以看到,即使是多类模式的分类,类间散度和识别率之间同样存在着非线性关系.

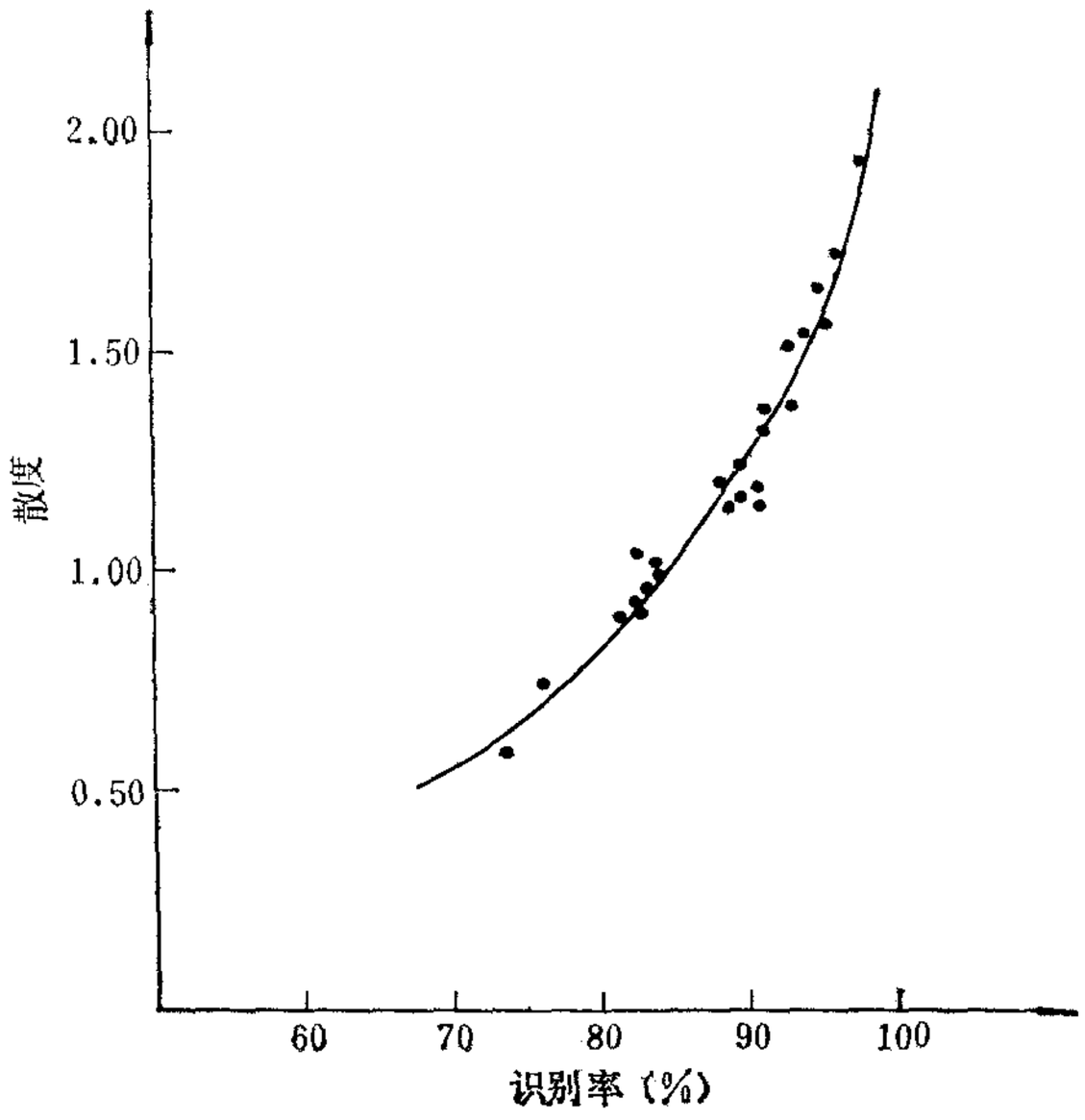

图 4 五类细胞散度与识别率关系图

在统计模式识别中,特征选择还有一些常规的方法,如逐步回归法,逐步判别法等.与这些方法相比较,本文介绍的特征选择法计算较简单,它克服了大数组运算的缺点,分类效果也比较好. 通过类间散度与识别率之间关系的分析,把模式识别过程中特征选择与

分类判别这两个阶段有机地联系起来，它对今后进一步开展这方面的研究工作将是有用的.

这项工作得到我所洪继光和有关同志的热情帮助和大力支持，在此表示衷心的感谢.

参　考　文　献

[1] 福永圭之介，统计图象识别导论，陶笃纯译，科学出版社，1978 年，246.
[2] Ashok, v. Kulkarni, Effectiveness of Feature Groups for Automated Pairwise Leucocyte Class Discrimination, *The Journal of Histochemistry and Cytochemistry*, (1979) 1, 210—216.
[3] 洪继光等，细胞图象的特征描述，信息与控制，1983 年第 12 卷第 2 期.
[4] 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，206—210.

A METHOD OF BETWEEN-CLASS DIVERGENCE FEATURE SELECTION AND ITS APPLICATION

QIAN XUESHUANG
(Institute of Automation, Academia Sinica)

In this paper, a method of feature selection in statistical pattern recognition the between-class divergence algorithm is introduced. It takes as its criterion of feature selection the ratio of between-class difference to within-class difference of the point set obtained from Fisher's mapping of two classes of cell smaple in n-dimentional feature space. The method is applied to feature selection for five classes of leucocyte, and the recognition effectiveness of the selected features is examined with train set and test set. The relation between feature selection criterion D and the correct recognition rale is obtained.