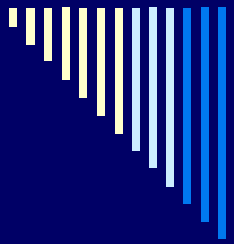
DNA序列的分类模型

一、问题

假定已知两组人工已分类的DNA序列（20个已知类别的人工制造的序列），其中序列标号1—10为A类，11-20为B类。要求我们从中提取已经分类了的DNA序列片段的特征和构造分类方法，并且还要衡量所用分类方法的好坏，从而构造或选择一种较好的分类方法。测试对象是20个未标明类别的人工序列（标号21—40）和182个自然DNA序列。例如A类：



```
a1='aggcacggaaaaacgggaataacggaggaggacttggcacggcattaca
cggaggacgaggtaaaggaggcttgtctacggccggaagtgaagggggatatg
accgcttgg';

b1='gttagatttaacgtttttttatggaatttatggaattataaatttaaaaa
tttatatttttttaggtaagtaatccaacgtttttattacttttttaaaattaaa
tatttatt';

......
```

二、特征提取

序列中含有四个碱基a、g、t、c，反映该序列特征的方面主要有两个：

1、碱基的含量，反映了该序列的内容；

统计a、g、t、c序列中分别出现的频率；

记序列中A、G、T、C的含量百分比为na、ng、nt、nc，则得到一组表征该序列特征的四维向量。

对于标号为i的序列，记它的特征向量为

$X_i = (na，ng，nt，nc)_i$。

$$X_A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & & & x_{24} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{bmatrix} \quad Y_B = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & & & y_{24} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & y_{n3} & y_{n4} \end{bmatrix}$$

统计出的数据结构为：

	na ng nt nc
1 2 … 10	

学习样本

A、B两类分别为：

A类的几何中心：$\mu_A = (\mu_{A1}\ \mu_{A2}\ \mu_{A3}\ \mu_{A4})$

B类的几何中心：$\mu_B = (\mu_{B1}\ \mu_{B2}\ \mu_{B3}\ \mu_{B4})$

欲判别类别的样本$X_i=(x_{i1},x_{i2},x_{i3},x_{i4}),i=21,\ldots,40;$

2、 碱基的排序

① 字符出现的周期性；

② 统计三个字符出现的频率；在遗传学中每三个碱基的组合被称为一个密码子，如agg，att，gag等，共有$4^3=64$个。其数据结构：

$$X_A=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,64} \\ x_{21} & & \cdots & x_{2,64} \\ \vdots & & \cdots & \vdots \\ x_{101} & x_{102} & \cdots & x_{10,64} \end{bmatrix} \quad Y_B=\begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,64} \\ y_{2,1} & & \cdots & y_{2,64} \\ \vdots & & \cdots & \vdots \\ y_{10,1} & y_{10,2} & \cdots & y_{10,64} \end{bmatrix}$$

三、建立分类模型

主要有三种分类模型：

- 统计分类模型

  距离判别、Fisher判别、Bayes判别等

- 建立信息量函数（熵函数）
- 神经网络模型

降维处理

如何将64个密码子减成几个？

建立的准则是 $|p_{Ai} - p_{Bi}| > 0.05$ 见表1

经分析知，可以将64维的密码子简化为只有8维的密码子。

A类序列的特征密码子：GGA，CGG，GGC，AGG

B类序列的特征密码子：TTA，TTT，ATT，TAT

模型一：构造辨析纸

1、计算特征密码子出现频率

八个密码子：

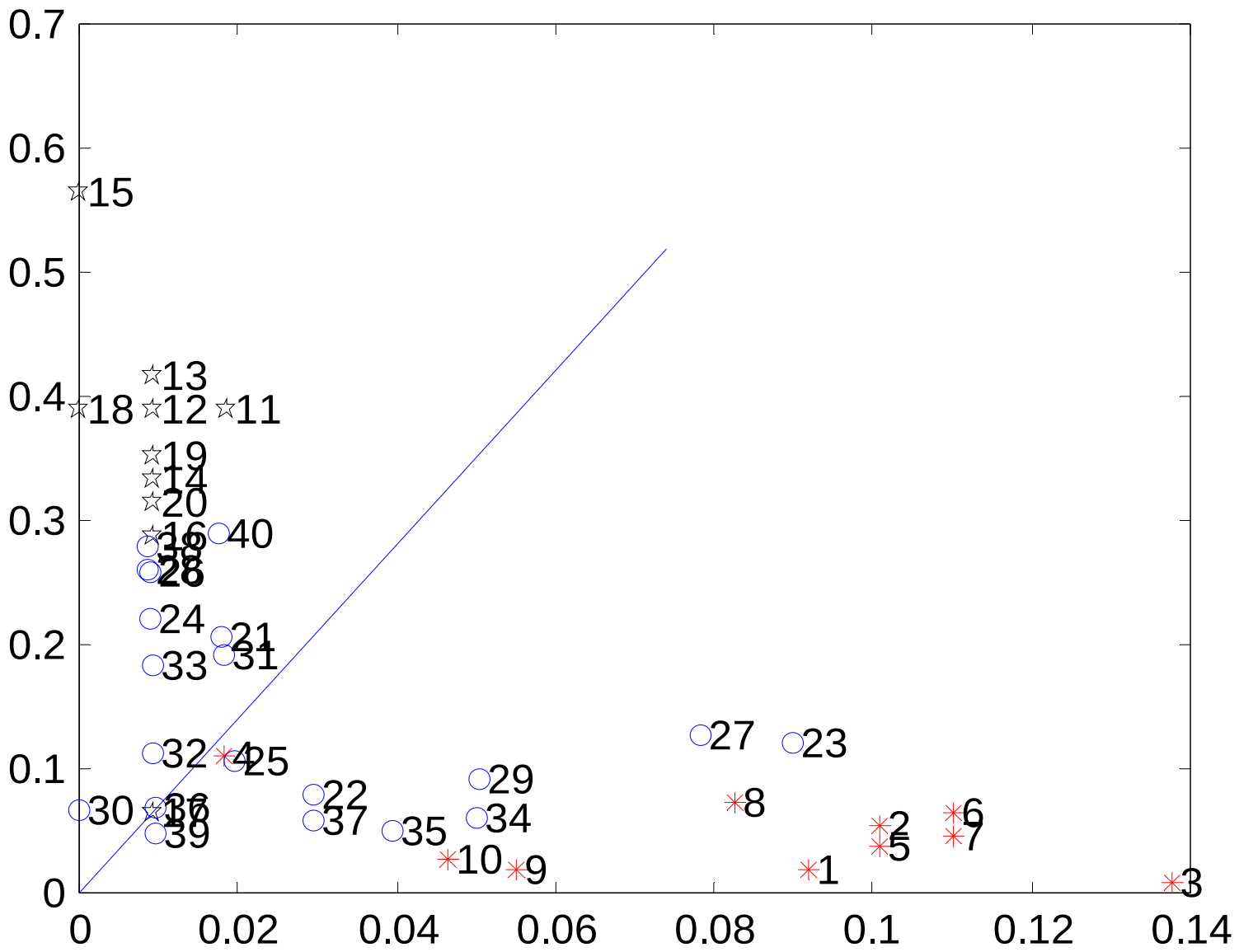
GGA，CGG，GGC，AGG，TTA，TTT，ATT，TAT，

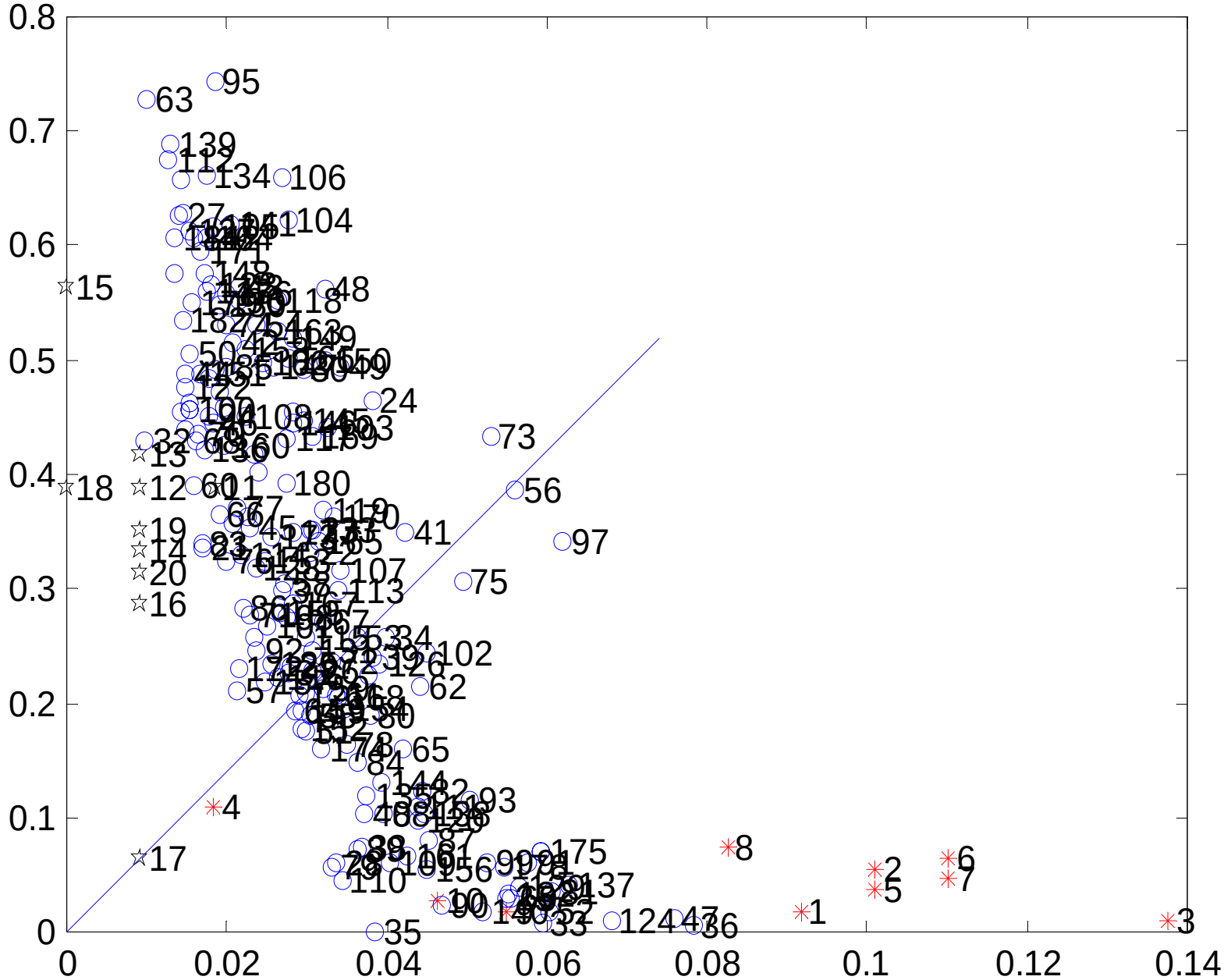
A　　B

其数据结构为

$$P_A = (\underbrace{p_{i1}, p_{i2}}_{p'_{i1},}, \cdots, \underbrace{p_{i7}, p_{i8}}_{p'_{i2}}), \quad i = 1, 2, ..., 10$$

$$P_B = (q_{i1}, q_{i2}, \cdots q_{i8}), \quad i = 1, 2, ..., 10$$

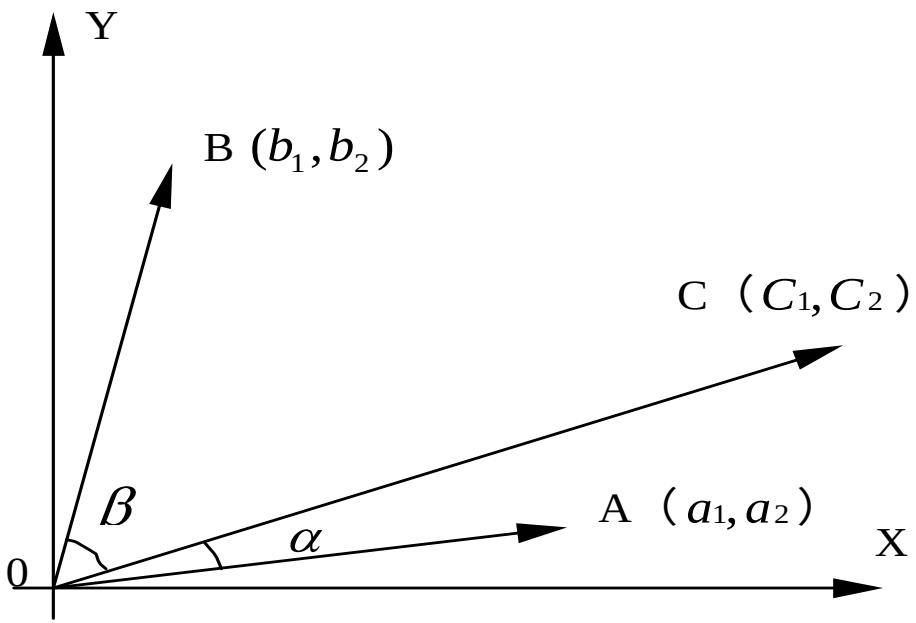

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
15
13
18
12
11
19
14
20
16
40
24
21
31
33
32
4
25
22
29
30
37
34
35
39
10
9
27
23
8
2
6
5
7
1
3


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14

对20个人工DNA序列进行分类，准确率已经达到95%。

对182个自然序列进行分类其准确率不高，必须采用其它方法进行分类。

模型二：多维向量空间的判别分析模型



图四　二维空间向量夹角示意图

如上图所示，向量OA、OB分别代表了A、B两类向量的重心位置。OC是任一个二维向量，与OA、OB成夹角α和β。

当 $\alpha < \beta$ 时 $\cos\alpha - \cos\beta < 0$ 可断定OC属于A类，否则，OC属于B类。’

$$\cos\alpha = \frac{a_1c_1 + a_2c_2}{\sqrt{a_1^2 + a_2^2}\cdot\sqrt{c_1^2 + c_2^2}}$$

$$\cos\beta = \frac{b_1c_1 + b_2c_2}{\sqrt{b_1^2 + b_2^2}\cdot\sqrt{c_1^2 + c_2^2}}$$

定义判别式：

$$W = \frac{a_1c_1 + a_2c_2}{\sqrt{a_1^2 + a_2^2}\cdot\sqrt{c_1^2 + c_2^2}} - \frac{b_1c_1 + b_2c_2}{\sqrt{b_1^2 + b_2^2}\cdot\sqrt{c_1^2 + c_2^2}}$$

判断准则如下：

1）当W>0时，判断向量OC属于A类；

2）当W<0时，判断向量OC属于B类；

3）当W=0时，不能判断；

将2维向量推广到64维向量，向量中的每个元素对应一个密码子在这个片段中出现的频率，第i个片段的向量表示为：

$$P_i = (p_{i,1}, p_{i,2}, \cdots, p_{i,64}), \quad i = 1, 2, ..., 10$$

而A、B两类的重心向量分别为：

$$\hat{P}_A = \frac{1}{10}\left[\sum_{m=1}^{10} p_m(1), \sum_{m=1}^{10} p_m(2), \ldots, \sum_{m=1}^{10} p_m(64)\right]$$

$$\hat{P}_B = \frac{1}{10}\left[\sum_{n=11}^{20} p_n(1), \sum_{n=11}^{20} p_n(2), \ldots, \sum_{n=11}^{20} p_n(64)\right]$$

由此可计算夹角余弦，从而计算判别函数： $W = \cos\alpha - \cos\beta$

$$\cos\alpha = \frac{\sum_{k=1}^{64} p_j(k)\hat{P}_A(k)}{\left[\sum_{k=1}^{64} p_j{}^2(k) \sum_{k=1}^{64} \hat{P}_A{}^2(k)\right]^{\frac{1}{2}}}$$

$$\cos\beta = \frac{\sum_{k=1}^{64} p_j(k)\hat{P}_B(k)}{\left[\sum_{k=1}^{64} p_j{}^2(k) \sum_{k=1}^{64} \hat{P}_B{}^2(k)\right]^{\frac{1}{2}}}$$

思考：

1、如何统计DNA序列片段中碱基a，g，t，c的频率；编程实现。

2、试分别用统计方法（欧氏距离、马氏距离和Fisher判别）对人工或自然序列进行分类。

3、DNA序列的特征提取其它方法。