模糊聚类分析程序及应用

黄勇，隋益虎

（安徽技术师范学院，凤阳 233100）

模糊聚类分析是依据客观事物间的特征、亲疏程度和相似性，通过建立模糊相似关系对客观事物进行分类的方法。在蔬菜育种中，利用模糊聚类分析可以对品种资源进行数量分类，为选择优良杂种一代亲本提供依据。一般情况下，双亲关系愈远，差别愈大，杂种优势愈强。蔬菜的许多经济性状（如甘蓝叶球的大小、紧密度、株高、株幅等）都是数量性状，把这几个性状综合起来考虑，单凭肉眼是很难准确地把它们划分成几类的。模糊聚类分析可把蔬菜区分为不同的类，然后根据类间亲本杂交的杂种优势强于类内亲本间杂交的杂种优势的原则选配亲本，再结合其它方法便可提高育种效率。模糊聚类分析要处理大量的数据，若用人工计算是很繁琐的，甚至是不可能的，作者根据模糊聚类分析（模糊等价矩阵聚类分析方法）的数学原理，设计了 FoxBASE⁺的应用程序，运行此程序，可方便、快捷地使用模糊聚类分析，解决蔬菜育种及类似问题。

1 模糊等价矩阵聚类分析方法的数学原理

在模糊聚类分析中，要进行分类的对象称为样本。设有 n 个样本，即被分类对象的集合为：$X=\{x_1, x_2, \ldots x_n\}$，每一个样本 x_i 有 m 个特性指标，即样本 x_i 可表示为特性指标向量度 $x_i=(x_{i1}, x_{i2}, \ldots x_{im})$，则 n 个样本的特性指标矩阵为：$[x_{ij}]_{n\times m}$。

由于 m 个特性指标的量纲和数量级都不相同，分类时缺乏统一的尺度，为了消除特性指标单位的差别和特性指标数量级不同的影响，可用下列式子对各指标值 X_{ij}（第 i 个样本的第 j 个特性）施行数据规格化，获得规格化后的值 x'_{ij}：

$$x'_{ij}=\frac{x_{ij}-\overline{x_j}}{\sigma_j}$$

$$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij},\qquad \sigma_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2}$$

对于样本空间 $X=[x_{ij}]_{n\times m}$，设 x_{ij} 均已规格化，用多元分析的方法来建立样本与样本之间的相似关系（亲疏关系），得到一个样本与样本之间的模糊相似关系矩阵：$[r_{ij}]_{n\times n}$。确定 r_{ij} 的工作叫标定，标定的方法很多，可选择用得较多的欧氏矩离法来标定：

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$$

$r_{ij}=1-cd_{ij}$ 式中 C 是适当选择的常数，为使 $0\le r_{ij}\ge 1$，c 可以选最大的 d_{ij} 的倒数。

一个模糊等价关系（模糊等价矩阵）可以确定一个模糊分类。通过标定所得的模糊相似关系矩阵是模糊相容矩阵，但未必是模糊等价矩阵。因此，要对样本空间 X 进行分类，必须由模糊相容矩阵构造出新的模糊等价矩阵。构造模糊等价矩阵可使用平方自合成法。R^* 计算出模糊等价矩阵后，选定适当的 λ 值（截距），对其进行截割即模糊聚类。聚类原则是：x_i 与 x_j 在 λ 水平上时属于同类，当 $r_{ij}\ge\lambda$ 时，x_i 与 x_j 归为一类。

2 程序

下面是模糊聚类分析程序 mhjl.prg，在汉字 FoxBASE⁺ 2.10 环境下运行。

```
set talk off
inpu "输入样本数 n:" to n
inpu "输入特性指标数 m:" to m
dime x(n,m)
i=1
do while i<=n
   j=1
   do while j<=m
```

```
      inpu "输入原始数据:" to x(i,j)
      j=j+1
   enddo
   i=i+1
enddo
dime ssj(m),ssj_(m)
j=1
do while j<=m
   s1=0
   i=1
   do while i<=n
      s1=s1+x(i,j)
      i=i+1
   enddo
   ssj(j)=s1/n
   s2=0
   i=1
   do while i<=n
      s2=s2+(x(i,j)-ssj(j))^2
      i=i+1
    enddo
    ssj_(j)=sqrt(s2/(n-1))
    j=j+1
enddo
i=1
do while i<=n
   j=1
   do while j<=m
      x(i,j)=(x(i,j)-ssj(j))/ssj_(j)
      j=j+1
    enddo
   i=i+1
enddo
dime r1(n,n)
max=0
i=1
do while i<=n
   j=1
   do while j<=n
      s=0
      k=1
      do while k<=m
         s=s+(x(i,k)-x(j,k))^2
         k=k+1
      enddo
      r1(i,j)=sqrt(s)
      if r1(i,j)>max
         max=r1(i,j)
      endif
      j=j+1
    enddo
    i=i+1
enddo
i=1
do while i<=n
   j=1
   do while j<=n
      r1(i,j)=1-r1(i,j)/max
      j=j+1
   enddo
   i=i+1
enddo
e=2
do while e<=n
   if e<10
      x="r"+str(e,1)
      dime &x.(n,n)
   else
      x="r"+str(e,2)
      dime &x.(n,n)
   endif
   e=e+1
enddo
f=1
c=0
do while f=1
   f=0
   c=c+1
   y="r"+str(c,1)
   z="r"+str(c+1,1)
i=1
do while i<=n
   j=1
   do while j<=n
      t=1
      max=0
      do while t<=n
         min=min(&y.(i,t),&y.(t,j))
         if min>=max
            max=min
         endif
         t=t+1
      enddo
      &z.(i,j)=max
      if &y.(i,j)<>&z.(i,j)
         f=1
      endif
      j=j+1
   enddo
   i=i+1
  enddo
enddo
i=1
do while i<=n
   j=1
   do while j<=n
      ?? str(&z.(i,j),4,2)
      ?? " "
      j=j+1
    enddo
```

```
  ?                                    set talk on
  i=i+1                                retu
enddo
```

3 应用举例

表1是某12个甘蓝品种9个性状的观测值，用模糊聚类分析方法对此分类。

表1 12个甘蓝品种9个性状的平均观测值（原始数据）

品种号	株高(cm)	开展度(cm)	外叶数	叶球重(g)	叶球纵径(cm)	叶球横径(cm)	球形指数	紧实度	中心柱长(cm)
1	28.4	60.3	9.8	0.73	12.8	17.6	0.73	0.35	5.6
2	29.2	62.4	12.6	0.96	13.2	20.7	0.64	0.33	6.9
3	35.5	67.0	12.9	1.50	15.0	23.4	0.64	0.35	11.2
4	30.2	64.0	10.8	0.95	12.6	20.0	0.63	0.36	7.6
5	35.9	75.6	13.0	1.38	14.5	21.3	0.68	0.40	6.9
6	21.2	43.7	11.0	1.07	11.4	16.0	0.72	0.70	6.6
7	21.6	41.6	18.2	0.75	11.3	14.8	0.76	0.58	6.5
8	31.0	57.7	11.0	1.27	18.2	22.6	0.80	0.26	7.1
9	26.2	50.9	19.8	0.87	10.3	15.6	0.66	0.66	6.8
10	27.4	60.1	8.8	1.79	12.6	22.2	0.57	0.55	7.7
11	32.3	61.6	17.0	0.80	11.1	14.8	0.75	0.63	7.4
12	30.8	65.3	16.4	0.91	13.3	17.6	0.76	0.42	8.0

首先启动UCDOS，再启动FoxBASE⁺，出现圆点状态后可打入下列命令运行该程序：

```
.do  mhjl
```

此时屏幕上出现“输入样本数n:”，使用者可输入参与分类的样本数：12（表1中甘蓝品种数为12）；接着屏幕提示：“输入特性指标数m:”，此时输入9（株高、开展度、……中心柱长共9个指标）；再接着屏幕提示：“输入原始数据:”此时可将表1中的原始数据按行依次逐一输入，即28.4,60.3,……,0.42,8.0。每输入一个数据，按一次回车。所有数据输完后，即得如下模糊等价矩阵。

$$\begin{pmatrix}
1.00 & 0.69 & 0.52 & 0.69 & 0.64 & 0.65 & 0.65 & 0.51 & 0.65 & 0.55 & 0.65 & 0.65 \\
0.69 & 1.00 & 0.52 & 0.88 & 0.64 & 0.65 & 0.65 & 0.51 & 0.65 & 0.55 & 0.65 & 0.65 \\
0.52 & 0.52 & 1.00 & 0.52 & 0.52 & 0.52 & 0.52 & 0.51 & 0.52 & 0.52 & 0.52 & 0.52 \\
0.69 & 0.88 & 0.52 & 1.00 & 0.64 & 0.65 & 0.65 & 0.51 & 0.65 & 0.55 & 0.65 & 0.65 \\
0.64 & 0.64 & 0.52 & 0.64 & 1.00 & 0.64 & 0.64 & 0.51 & 0.64 & 0.55 & 0.64 & 0.64 \\
0.65 & 0.65 & 0.52 & 0.65 & 0.64 & 1.00 & 0.65 & 0.51 & 0.65 & 0.55 & 0.65 & 0.65 \\
0.65 & 0.65 & 0.52 & 0.65 & 0.64 & 0.65 & 1.00 & 0.51 & 0.69 & 0.55 & 0.66 & 0.66 \\
0.51 & 0.51 & 0.51 & 0.51 & 0.51 & 0.51 & 0.51 & 1.00 & 0.51 & 0.51 & 0.51 & 0.51 \\
0.65 & 0.65 & 0.52 & 0.65 & 0.64 & 0.65 & 0.69 & 0.51 & 1.00 & 0.55 & 0.66 & 0.66 \\
0.55 & 0.55 & 0.52 & 0.55 & 0.55 & 0.55 & 0.55 & 0.51 & 0.55 & 1.00 & 0.55 & 0.55 \\
0.65 & 0.65 & 0.52 & 0.65 & 0.64 & 0.65 & 0.66 & 0.51 & 0.66 & 0.55 & 1.00 & 0.70 \\
0.65 & 0.65 & 0.52 & 0.65 & 0.64 & 0.65 & 0.66 & 0.51 & 0.66 & 0.55 & 0.70 & 1.00
\end{pmatrix}$$

根据上述模糊等价矩阵中数值的大小，可对12个甘蓝品种进行分类。截矩λ分别选定为0.69,0.65,0.64,0.55,0.51,分类结果是：

{1,2,4}　{6,7,9,11,12}　{5}　{10}　{3,8}

由此得出分析结论：12个甘蓝品种可分为5类。1，2，4三个品种差异小归为第一类；6，7，9，11，12五个品种差异也小，归为第二类；……;3和8两者之间差异不大归为第五类。第五类与其它的品种差异较大，与第一类差异最大。因此，育种时应首先考虑在1，2，4三个品种中选择一个父本或母本，在3，8中选择一个母本或父本，这样的亲本组配得到F1杂种优势最强的概率最大，即最有希望选育出优良的杂交新品种来。（参考文献略）