第 9 章　判别分析

判别分析是一种常用的统计分析方法。判别分析是根据观察或测量到若干变量值，判断研究对象如何分类的方法。例如，我们积累了某种病虫害各种发生状态的若干历史资料样本)，希望从中总结出分类的规律性(即判别公式，在以后的工作中遇到新的发生状态(样本)时。只要根据总结出来的判别公式判断它所属的类就行了。动物、植物分类等都可以用判别分析来解决。

进行判别分析必须已知观测对象的分类和若干表明观测对象特征的变量值。判别分析就是要从中筛选出能提供较多信息的变量并建立判别函数，使得利用推导出的判别函数对观测量判别其所属类别时的错判率最小。

判别函数一般形式是：

$$Y = a_1X_1+a_2X_2+a_3X_3 \cdots +a_nX_n$$

其中:Y 为判别分数(判别值)；X_1，X_2，X_3：…X_n 为反映研究对象特征的变量，a_1、a_2、a_3…a_n 为各变量的系数，也称判别系数。可以看出我们这里所讲的是线性判别函数。

SPSS 对于分为 m 类的研究对象，建立 m 个线性判别函数。对于每个个体进行判别时，把测试的各变量值代入判别函数，得出判别分数，从而确定该个体属于哪一类。或者计算属于各类的概率，从而判断该个体属于哪一类。还可建立标准化和未标准化的典则判别函数。

SPSS 提供的判别分析过程是 Discriminant 过程。

[例子 9-1]

表 9-1 浙江北部地区 1950~1982 年小麦赤霉病发生程度与气象因子数据表

X1	X2	X3	X4	X5	y
14.3	107.3	140.0	105.3	6.9	1
46.5	129.1	154.1	91.3	11.9	1
43.0	143.1	83.9	157.4	13.0	2
71.2	280.5	82.5	317.4	13.9	3
.7	69.3	145.6	69.5	11.3	1
123.9	297.3	64.6	307.2	13.7	3
85.4	115.4	39.4	144.7	11.1	1
38.4	77.3	94.6	143.2	13.9	2
79.6	96.8	85.4	99.0	9.6	2
33.4	74.7	129.5	103.4	9.9	1
48.1	95.9	155.3	92.0	10.5	1
7.7	116.3	158.2	148.1	15.1	1
8.9	225.3	104.2	195.5	13.8	1
34.8	150.7	165.0	124.6	11.9	1
44.4	147.2	88.3	158.7	12.7	2
74.2	232.7	94.1	154.6	13.5	3
.1	80.9	148.8	81.3	11.0	1
119.6	208.0	70.9	217.8	13.8	3
94.0	130.2	49.2	176.2	11.0	2
32.9	83.6	115.3	135.7	13.8	2
65.5	88.1	126.9	102.5	9.7	1
31.3	59.3	105.1	82.9	10.0	1
52.3	93.3	173.7	91.2	10.0	1
7.2	98.2	154.3	120.7	15.0	1
5.3	245.8	100.4	200.2	13.7	1

浙江北部地区 1950~1982 年小麦赤霉病发生程度与气象因子研究，总结出上年 12 月将与（x1）、上年 10 月下旬至 11 月中旬和当年 1~2 月总降雨（x2）、上年 10 月下旬至 11 月上旬日照时数（x3）、上年 10 月下旬至 12 月中旬和当年 2 月总雨量（x4）以及当年 3 月中旬平均高文（x5）等 5 个因子，并将赤霉病情分为轻中重三级（y，分别用 1、2、3 表示）。数据见表 9-11。用这些数据建立气象因子与小麦赤霉病发生程度的判别模型。

9.1 操作方法

1）数据准备

在数据管理窗口，定义变量名 x1、x2、x3、x4、x5、y 分别表示表中对应变量。然后输入对应的数据。数据保存在配套光盘中（\SPSS\DATA\DATA9-1.SAV）。

2）启动分层聚类过程

在 SPSS 主菜单中按“Analyze→Classify→Discriminant”顺序逐一单击鼠标键，打开判别分析主对话框，如图 9-1 所示。

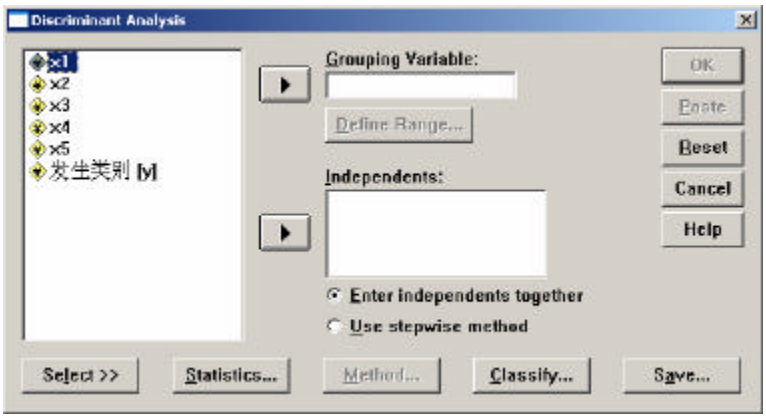
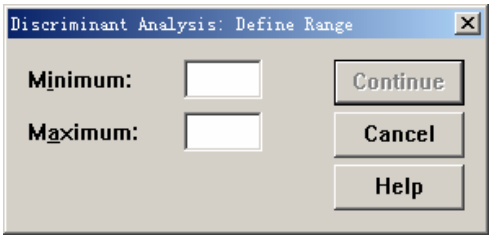


图 9-1 判别分析主对话框

3) 指定分析变量

指定分类变量及其范围： 在主对话框中左边的矩形框中选择已知的观测量所属类别的变量（一定是离散变量），点击“Grouping Variable”框左边的右拉箭头按钮，使该变量名移到该矩形框中。此时“Grouping Variable”矩形框下面的“Define Range”按钮加亮，按该按钮，屏幕显示一个小对话框，提供指定该分类变量的数值范围，如图 9-2 所示。



如图 9-2 定义分类变量范围对话框

Minimum 栏输入最小值。

Maximum 栏输入最大值。

本例选择“发生类别[y]”到“Grouping Variable”框中；在图 9-2 中的“Minimum 栏”输入最小值 1，“Maximum” 栏输入最大值 3。

指定判别分析的自变量：在主对话框中左边的变量框中选择表明观测量特征的变量，点击“Independents”框左边的右拉箭头按钮，使该变量名移到该矩形框中，作为参与判别分析的自变量。

本离例子选择“x1、x2、x3、x4、x5”变量到“Independents”矩形框的变量列表中。

选择进入分析的观测量：如果希望使用一部分观测量进行判别函数的推导，而且有一个变量的某个值可以作为这些观测量的标识，单击‘Selection’按钮，展开‘Selection”选择框，并从变量列表框中选中该变量；再单击“Selection”选择框右侧的“Valve”按钮，展开“Set Value”(子对话框)对话框，键入标识参与分析的观测量所具有的该变量值。

一般均使用数据文件中的所有合法观测量。此步骤可以省略。

4）选择分析方法

在主对话框中自变量矩形框下面有两个选择项选择判别分析方法：

- Enter independent together 选项。当认为所有自变量都能对观测量特性提供丰富的信息时使用该选择项。建立全模型。系统缺省设置。
- Use stepwise method 进入判别模型的自变量根据对判别贡献的大小进行逐步选择。选中该项后，“Method”按钮加亮。可以进一步选择判别分析方法（见下一步）。

本例选“Use stepwise method”项，进行逐步选择判别分析。

5) 设置逐步判别分析

在主对话框中单击“Method”按钮，打开设置逐步判别分析方法对话框，见图 9-3。

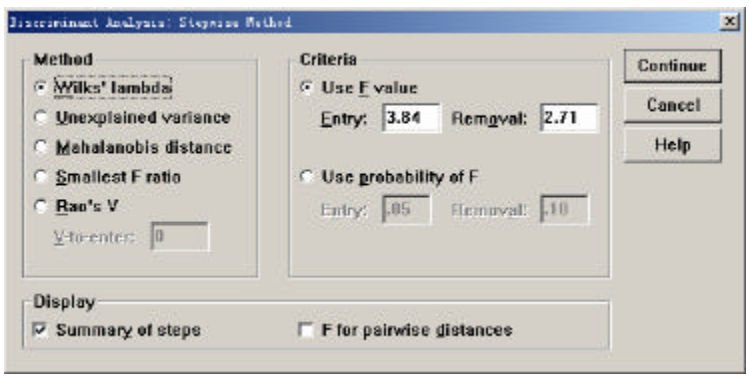


图 9-3 设置逐步判别分析方法

① “ Method ”栏选项选择:

- Wilks′ lambda 选项，每步都是 Wilk 的入统计量最小的进入判别函数。
- Unexplained variance 选项，每步都是使各类不可解释的方差和最小的变量进入判别函数。
- Mahalanobis distance 选项，每步都使靠得最近的两类间的 Mahalanobis 距离最大的变量进入判别函数。
- Smallest F ratio 选项，每步都使任何两类间的最小的 F 值最大的变量进入判别函数。
- Rao′s V 选项，每步都是使 Rao′s V 统计量产生最大增量的变量进入判别函数。可以对一个要加入到模型中的变量的 V 值指定一个最小增量。选择此种方法后，应该在该项下面的 V-to-enter 后的矩形框中输入这个增量的指定值。当某变量导致的 V 值增量大于指定值的变量进入判别函数。

本例选中“ Mahalanobis distance ”选项。

② “ Criteria ”选择逐步判别停止的判据

选择逐步判别停止的判据在“ Criteria ”栏中进行。可供选择的判据有：

- Use F value 选项，使用 F 值，是系统默认的判据，当加入一个变量(或剔除一个变量)后，对在判别函数中的变量进行方差分析。当计算的 F 值大于指定的 Entry 值时，该变量保留在函数中。默认值是 Entry 为 3.84；当该变量使计算的 F 值小于指定的 Removal 值时，该变量从函数中剔除。默认值是 Removal 为 2.71。即当被加入的变量 F 值≥3.84 时才把该变量加入到模型中，否则变量不能进入模型；或者当要从模型中移出的变量 F 值≤2.71 时，该变量才被移出模型，否则模型中的变量不会被移出。设置这两个值时应该注 Entry 值大于 Removal 值。
- Use probability of F 选项，用 F 检验的概率决定变量是否加入函数或被剔除而不是用 F 值。加入变量的 F 值概率的默认值是 0.05(5%)；移出变量的 F 值概率是 0.10(10%)。Removal 值(移出变量的 F 值概率)大于 Entry 值(加入变量的 F 值概率)。

本例选中“ Use F value ”选项，“ Entry ”栏输入 3.0，“ Removal ”输入 2.0

③ “ Display” 显示内容的选择

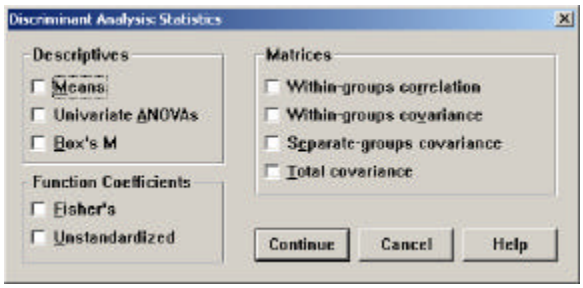
对于逐步选择变量的过程和最后结果的显示，可以通过“ Display ”栏中的两项进行选择：

- Summary of steps 复选项，要求在逐步选择变量过程中的每一步之后显示每个变量的统计量。
- F for Pairwise distances 复选项，要求显示两类之间的 F 值矩阵。

本例两项都不选择。

6）统计量输出设置

在主对话框中点击“Statistic”按钮，打开统计量输出设置对话框，如图 9-4。



如图 9-4 “Statistic”对话框

① “Descriptives”栏选择输出描述统计量:

- Means 复选项，可以输出各类中各自变量的均值 Mean、标准差 Std.Dev 和各自变量总样本的均值和标准差。
- Univarlate ANOVAs 复选项，对各个自变量进行均值假设检验，输出单变量的方差分析结果。
- Box’s M 复选项，对各类的协方差矩阵相等的假设进行检验。

本例选中“Means”选项。

②“Function coefficients”栏选择输出判别函数系数

- Fisher’s 复选项，可以直接用于对新样本进行判别分类的费雪系数。对每一类给出一组系数。并给出该组中判别分数最大的观测量。
- Unstandardized 复选项，未经标推化处理的判别系数。

本例选中“Fisher’s”选项。

③“Matrices”栏选择输出自变量的系数矩阵

- Within-groups correlation matrix 复选项，即类内相关矩阵，它是根据在计算相关矩阵之前将各组(类)协方差矩阵平均后计算类内相关矩阵。
- Within-groups covariance matrix 复选项，即计算并显示合并类内协方差矩阵，是将各组(类)协方差矩阵平均后计算的。区别于总协方差阵。
- Separate-groups covariance matrices 复选项，对每类输出显示一个协方差矩阵。
- Total covariance matrix 复选项，计算并显示总样本的协方差矩阵。

本例子 4 项都不选择。

7）“Classify”栏指定分类参数和判别结果

在主对话框中单击“Classify”按钮，打开分类参数设置对话框，如图 9-5 所示。

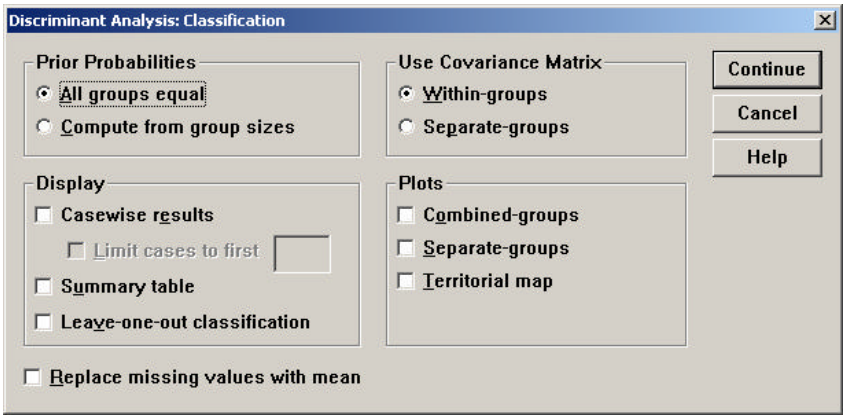


图 9-5 分类参数设置对话框

① Prior Probabilities 栏中选择先验概率

- All groups equal 选项，各类先验概率相等。若分为 m 类，则各类先验概率均为 l／m。
- Compute from grous sizes 选项，由各类的样本量计算决定，即各类的先验概率与其样本量成正比。

本例子选中“Compute from grous sizes”选项。

②“Use Covariance Matrix”栏选择分类使用的协方差矩阵

- Within-groups 选项，指定使用合并组内协方差矩阵进行分类。
- Separate-groups s 选项，指定使用各组协方差矩阵进行分类。由于分类是根据判别函数而不是根据原始变量。因此该选择项不是总等价于二次判别。

本例选中“Within-groups”选项。

③“Display”栏选择生成到输出窗中的分类结果

- Casewise results 复选项，输出每个观测量包括判别分数、实际类、预测类(根据判别函数求得的分类结果)和后验概率等。选择此项，还可以选择其附属选择项：
  - ✧ Limits cases to 复选项，并在后面的小矩形框中输入观测量数 n。选择此项则仅输出前 n 个观测量。观测数量大时可以选择此项。
- Summary table 复选项，要求输出分类的小结，给出正确分类观测量数(原始类和根据判别函数计算的预测类相同)和错分观测量数和错分率。
- Leave-one-out classification 复选项，输出对每个观测量进行分类的结果，所依据的判别函数是由除该观测量以外的其他观测量导出的。也称为交互校验结果。

本例选中“Summary table”选项。

④“Plots”栏选择要求输出的统计图

- Combined-groups 复选项，生成一张包括各类的散点图。该散点图是根据前两个判别函数值作的散点图。如果只有一个判别函数，就输出直方图。
- Separate-groups 复选项，根据前两个判别函数值对每一类生成一张散点图。共分为几类就生成几张散点图。如果只有一个判别函数，就输出直方图。

- Territorial map 复选项，生成用于根据函数值把观测量分到各组中去的边界图。此种统计图把一张图的平面划分出与类数相同的区域。每一类占据一个区。各类的均值在各区中用星号标出。如果仅有一个判别函数，则不作此图。

本例不输出图形，3 项都不选。

⑤ 缺失值处理方式

- Replace missing value with mean 复选项，即用该变量的均值代替缺失值。

8）选择要存入数据文件的新变量

在主对话框中单击“ Save ”按钮，打开“ Save ”对话框，如图 9-6 所示。

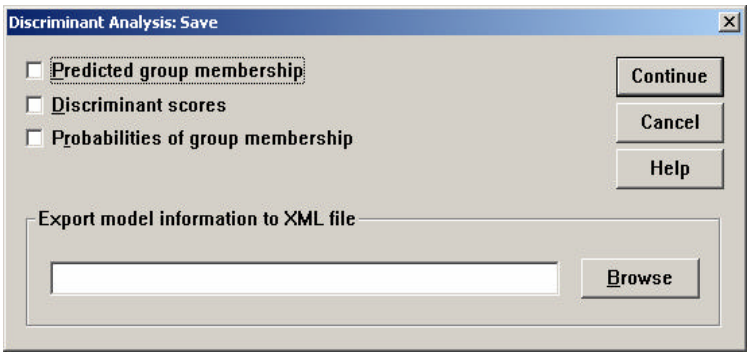


图 9-6 “ Save ”对话框

- Predicted group membership 复选项，要求建立一个新变量，表明观测量被预测的分类。是根据判别分数把观测量按后验概率最大指派所属的类，每运行一次判别过程，就建立一个表明使用判别函数预测的各观测量属于哪一类的新变量。第一次运行建立新交量的变量名为 dis_1，如果在工作数据文件中不把前一次建立的新变量删除第 n 次运行判别过程建立的新变量默认的变量名为 dis_n。
- Discriminant score 复选项，要求建立表明判别分数的新变量。该分数是由未标准化的判别系数乘自变量的值，将这些乘积求和后加上常数得来。每次运行判别过程都给出一组表明判别分数的新变量。建立几个判别函数就有几个判别分数变量。参与分析的观侧量共分为 m 类，则建立 m—1 个典则判别函数，指定该选择项，就可以生成 m—1 个表明判别分数的新变量。例如原始数据观测量共分为 3 类，建立两个典则判别函数。第一次运行判别过程建立的新变量名为 disl_1、dis2_l，第二次运行判别过程建立的新变量名为 disl_2、dls2_2…依此类推。分别表示代入第一和第二个判别函数所得到的判别分数。
- Probabilities of group membership 复选项，要求建立新变量表明观测量属于某一类的概率。有 m 类，对一个观测量就会给出 m 个概率值，因此建立 m 个新变量。例如原始和预测分类数是 3，指定该选择项，在第一次运行判别过程后，给出的表明分类概率的新变量名为 dis1_2、dis2_2、dis3_2。

本例选中“ Predicted group membership ”选项。

9）提交设置执行过程

各种选择项确定之后返回到主对话框，点击“OK”按钮，SPSS 输出结果将显示在输出浏览器和数据编辑窗口的工作文件中。

9.2 结果解释

表 9-2 **分组统计** Group Statistics

发生类别		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
1	X1	28.009	23.956	23	23.000
	X2	109.883	50.500	23	23.000
	X3	135.004	30.132	23	23.000
	X4	109.787	36.531	23	23.000
	X5	11.317	2.243	23	23.000
2	X1	53.500	22.522	9	9.000
	X2	108.878	30.023	9	9.000
	X3	86.278	19.053	9	9.000
	X4	140.611	23.097	9	9.000
	X5	12.400	1.518	9	9.000
3	X1	98.583	28.478	6	6.000
	X2	248.400	34.180	6	6.000
	X3	75.850	10.896	6	6.000
	X4	252.800	62.164	6	6.000
	X5	13.717	.147	6	6.000
Total	X1	45.189	34.973	38	38.000
	X2	131.516	67.083	38	38.000
	X3	114.124	36.491	38	38.000
	X4	139.668	63.732	38	38.000
	X5	11.953	2.073	38	38.000

表 9-2 **给出了数据分组后各组的均值、标准差，“** Valid N **”栏中选择了“** Unweighted **”项为未加权的观测量数目，“** Weighted **”项为已加权的观测量数目（每个观测量的权数为** 1**）。**

表9-3 **典则判别函数的特征值** Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.142[a]	79.6	79.6	.871
2	.807[a]	20.4	100.0	.668

a First 2 canonical discriminant functions were used in the analysis.

表9-4 **λ值的卡方转换及卡方检验** Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.134	67.422	8	.000
2	.554	19.812	3	.000

表9-5 标准化的典则判别函数系数表

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
X1	.842	.101
X2	.626	.818
X3	.018	.986
X5	.367	-.451

表9-6 结构矩阵 Structure Matrix

	Function	
	1	2
X2	.646*	.344
X1	.604*	-.207
X4[a]	.578*	.056
X5	.261*	-.145
X3	-.479	.684*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

* Largest absolute correlation between each variable and any discriminant function

a This variable not used in the analysis.

表9-7 未标准化的典则判别函数系数表

Canonical Discriminant Function Coefficients

	Function	
	1	2
X1	.035	.004
X2	.014	.018
X3	.001	.038
X5	.191	-.235
(Constant)	-5.775	-4.145

Unstandardized coefficients

表9-8 各类中心未标准化的判别函数系数表

Functions at Group Centroids

	Function	
发生类别	1	2
1	-1.005	.474
2	3.458E-02	-1.547
3	3.802	.502

Unstandardized canonical discriminant functions evaluated at group means

表9-9 线性判别的数据信息 Classification Processing Summary

Processed		38
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		38

表 9-10 各函数内自变量的先验概率 **Prior Probabilities for Groups**

发生类别	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1	.605	23	23.000
2	.237	9	9.000
3	.158	6	6.000
Total	1.000	38	38.000

表 9-11 **用于对数据分类的** Fisher **线性判别函数系数表**

Classification Function Coefficients

	发生类别		
	1	2	3
X1	.336	.364	.503
X2	7.728E-02	5.472E-02	.145
X3	.351	.275	.356
X5	3.917	4.590	4.828
(Constant)	-55.325	-54.470	-91.280

Fisher's linear discriminant functions

表 9-11 **是逐步判别后给出的判别函数系数表，最后进入的自变量是** x1、x2、x3 **和** x5**，其判别函数如下：**

$F1 = 0.336X_1 + 0.07728X_2 + 0.351X_3 + 3.917X_5 - 55.325$

$F2 = 0.364X_1 + 0.05472X_2 + 0.275X_3 + 4.590X_5 – 54.470$

$F3 = 0.503X_1 + 0.145X_2 + 0.356X_3 + 4.828X_5 – 91.28$

表 9-12 **判别回代统计表** **Classification Results**[a]

			Predicted Group Membership			Total
		发生类别	1	2	3	
Original	Count	1	22	1	0	23
		2	0	9	0	9
		3	0	0	6	6
	%	1	95.7	4.3	.0	100.0
		2	.0	100.0	.0	100.0
		3	.0	.0	100.0	100.0

a 97.4% of original grouped cases correctly classified.

表 9-12 **表明：赤霉病发生较轻（**1 **类）共有** 23 **年，用判别函数回代分类，与实际相符的** 22 **年，错分为一般发生（**2 **类）为** 1 **年，**1 **类的准确率为** 95.7%**。赤霉病一般发生（**2 **类）共有** 9 **年，用判别函数回代分类，与实际相符的** 9 **年，没有错分，**2 **类判别的准确率为** 100%**。赤霉病发生严重（**3 **类）共有** 6 **年，用判别函数回代分类，与实际相符** 6 **年，没有错分，因此** 3 **类判别的准确率为** 100%**。由此可以说明判别函数可信。**

逐步判别分析应该给出尽可能最佳的判别分析结果。判别计算结束后，如果判别效果不显著，或者正确判别率太低，则应该考虑改变挑选变量的统计量 F **值。降低或者提高** F **值，强制性的增加或减少选入判别函数的变量数。如果这种尝试能取得较高的正确判别率，则可作为最终的判别分析结果。如果仍不能获得较高的正确判别率，则只能选取一个相对较好的结果作为判别分析结果。**