

常用费歇判别准则的比较*

周静芋, 宋世德, 郭满才

(西北农林科技大学 生命科学学院, 陕西 杨陵 712100)

[摘要] 对费歇判别分析中用于判别样品归属的几个常用判别准则进行了讨论和比较, 认为用临界值准则判别的正确率较小, 即使用改进的临界值准则也未必能得到较大的改善, 而利用马氏距离法一般可得到更好的判别效果。在分析的基础上, 提出了一种新的马氏距离法。

[关键词] 费歇(Fisher)判别; 判别准则; 马氏距离

[中图分类号] O 212 1 [文献标识码] A

[文章编号] 1671-9387(2002)05-0121-03

费歇判别分析是运用方差分析的思想, 利用已知样本的分类, 用多元函数极值思想, 通过使类间方差达到最大而类内方差达到最小, 来求判别函数的一种判别分析方法。建立好判别函数后, 需要一定的判别准则来判断样品的归属, 常用的准则有临界值法和马氏距离法两种^[1~3]。在实际应用中, 由于人们对这两种方法的原理及其适用情况了解不够, 一般教科书上也很少提及, 常常导致一些应用上的混乱。本研究拟就这两种方法进行讨论和比较, 指出各自的优缺点, 以期能对其正确应用提供一定的参考, 同时提出一种新马氏距离法。

1 常用费歇判别准则

1.1 费歇判别函数

费歇判别思想是将各类(组)数据投影到某一个方向, 使得各类数据的投影尽可能地分开。若从 k 类分别取得 m 维观察值: $G_1: X_1^{(1)}, X_2^{(1)}, \dots, X_{n_1}^{(1)}, \dots; G_k: X_1^{(k)}, X_2^{(k)}, \dots, X_{n_k}^{(k)}$ 。令 $y(X) = a^T X$ 为 X 以 a 为法向方向的投影。费歇判别分析借用方差分析的思想来判别类与类之间是否尽可能地分开, 即要使每个类中 y 值离散程度越小越好, 不同类之间的 y 值离散程度越大越好。经数学推导求 a 归结为一个条件极值:

$$\begin{cases} a^T \hat{\Sigma} a = 1 \\ a^T B a = \max \end{cases}$$

式中, $\hat{\Sigma}$ 为类内协方差阵的无偏估计, $\hat{\Sigma} = \frac{1}{(n-k)} L_{XX}$ (在建立判别函数时, 本文仅讨论这种情况); B, L_{XX} 分别为类间和类内的离差阵。若 $\hat{\Sigma}$ 可逆, 可求得上式

的非零特征根为: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r (r \leq m)$, 相应的特征向量为: $a_1, a_2, \dots, a_r$, 由此得费歇判别函数:

$$y_t = a_t^T X, \quad t = 1, 2, \dots, r \quad (1)$$

第一个判别函数 y_1 的判别效率最大, 当只用一个判别函数时, 常将 y_1 记作 y 。各判别函数之间是互不相关的。

1.2 常用判别准则比较

当得到费歇判别函数后, 判断样品归属常用的方法有:

(1) 临界值法: 计算每一类的 $\bar{y}^{(i)} = a^T \bar{X}^{(i)}$ 值。将 $\bar{y}^{(1)}, \bar{y}^{(2)}, \dots, \bar{y}^{(k)}$ 由小到大重新排序为:

$$\bar{y}_*^{(1)} \leq \bar{y}_*^{(2)} \leq \dots \leq \bar{y}_*^{(k)} \quad (2)$$

此时将类的原编号重新编号, 将 k 个平均值每相邻 2 个取加权平均, 即

$$\bar{y}_c^{(i)} = \frac{1}{n_i + n_{i+1}} (n_i \bar{y}_*^{(i)} + n_{i+1} \bar{y}_*^{(i+1)}) \quad (3)$$

式中, $i = 1, 2, \dots, k-1$, 用 $\bar{y}_c^{(1)}, \bar{y}_c^{(2)}, \dots, \bar{y}_c^{(k-1)}$ 作为 $y = a^T X$ 判别的临界值, 若一个样品 X 满足 $\bar{y}_*^{(i)} < y < \bar{y}_c^{(i+1)}$, 当 $y \geq \bar{y}_c^{(i)}$ 判定 X 属于第 $i+1$ 类; 当 $y < \bar{y}_c^{(i)}$, 则判定 X 属于第 i 类。

这种方法的优点是当第一个判别函数 $y_1 = a_1^T X$ 的效果(指判别能力或效率, 它等于 $\frac{\lambda_1}{\sum \lambda_j}$ 或回判正确率)较好时, 该方法简便直观, 但当各类 n_i 或 $\hat{\sigma}_i$ 相差较大时, 方法显得粗糙, 易造成较大偏差。因此又有改进方法:

$$\bar{y}_c^{(i)} = \frac{\hat{\sigma}_{i+1} \bar{y}_*^{(i)} + \hat{\sigma}_i \bar{y}_*^{(i+1)}}{\hat{\sigma}_i + \hat{\sigma}_{i+1}} \quad (4)$$

* [收稿日期] 2001-09-07

[作者简介] 周静芋(1944-), 女, 浙江鄞县人, 教授, 主要从事应用数学研究。

式中, $\hat{\sigma}_i$ 为相应新排类别判别函数值的标准差。这种方法有益于解决 n_i 或 $\hat{\sigma}_i$ 相差大的问题, 但仅适用于第一个判别函数判别效果好的情况, 当要用多个判别函数时就不方便了。

(2) 马氏距离法。计算任一样品 X 的判别函数值 $y_1 = a_1^T X$, 求它与第 i 个总体的马氏距离:

$$d_i^2 = (a_1^T X - a_1^T \hat{\mu})^T (a_1^T \hat{\Sigma} a_1)^{-1} (a_1^T X - a_1^T \hat{\mu}) = (X - \hat{\mu})^T a_1 (a_1^T \hat{\Sigma} a_1)^{-1} a_1^T (X - \hat{\mu}) \quad (5)$$

若 $d_i^2 = \min_{1 \leq i \leq k} \{d_i^2\}$, 则判 $X \in G_i$ 。

若用 $Y_1 = a_1^T X$ 判别的效率很低, 可相继计算 $Y_2 = a_2^T X, Y_3 = a_3^T X, \dots, Y_s = a_s^T X (s \leq r)$ 。则判别函数为: $w(X) = (a_1, a_2, \dots, a_s)^T X = A^T X, A = (a_1, a_2, \dots, a_s)$ 。对于样品 X 求 $w(X)$ 与各总体的马氏距离:

$$d_i^2 = (X - \hat{\mu})^T A (A^T \hat{\Sigma} A)^{-1} A^T (X - \hat{\mu}) \quad (6)$$

则有如下判别规则: 若 $d_i^2 = \min_{1 \leq i \leq k} \{d_i^2\}$, 则判 $BZ \in G_i$ 。其中各总体(各类)均值 μ_i 估计值为: $\hat{\mu} = \bar{X}^{(i)}$, 协方差阵 Σ 估计值为: $\hat{\Sigma} = \frac{1}{n_i - 1} L_{XX}^{(i)}$ 。

这种方法适用于要用多于一个判别函数的情况, 但计算量大, 在实际中较少使用, 常用的是:

$$d_i^2 = \sum_{j=1}^s (a_j^T X - a_j^T \bar{X}^{(i)})^2 = \sum_{j=1}^s (y_{ij} - \bar{y}_j^{(i)})^2 \quad (s < m) \quad (7)$$

这种方法计算 d_i^2 较用(6)简单得多, 但它的前提是使用合并组内协方差矩阵对样品归类, 即公式(6)中 $\hat{\Sigma}$ 皆用 $\frac{1}{n-k} L_{XX}$ 代替(此时 $A^T \hat{\Sigma} A = 1$)。这种方法用得最为广泛, 这时的马氏距离即为欧氏距离。考虑到每个判别函数的判别能力不同, 有时也用 d_i^2

$$= \sum_{j=1}^s (y_{ij} - \bar{y}_j^{(i)})^2 \lambda_{ij}$$

1.3 新马氏距离

本研究提出的不同于上面所介绍的两种马氏距离法的新马氏距离法, 其根本思想是基于按各类协方差矩阵进行分类的马氏距离法, 但不同于式(5)和(6)的是, 这里各类的协方差不是原始数据阵的协方差阵, 而是各类样品判别函数得分值的协方差阵。它的步骤是: 按前面所述求出 n 个样品的 s 个判别函数得分值, 然后用它们作为新的变量求出各类的协方差阵 $S^{(i)}$, 最后计算

$$d_i^2 = (X - \bar{X}^{(i)})^T A (S^{(i)})^{-1} A^T (X - \bar{X}^{(i)}) = (y - \bar{y}^{(i)})^T (S^{(i)})^{-1} (y - \bar{y}^{(i)}) \quad (8)$$

按 d_i^2 的大小进行分类, 当第一个判别函数判别

效果好时, (8) 式中的 A 可用 a_1 代替。易证 $S^{(i)} = A^T \hat{\Sigma} A$, 所以, 理论上用此马氏距离和用公式(6)计算的马氏距离是一样的。

2 实 例

拟对陕西省进行喷灌区划, 其一级区预分3类, 为说明判别的方法, 这里仅选出3个比较重要的变量, 并以从陕南、关中、陕北地区选择的27个变量作为样本(第1类15个, 第2类3个, 第3类9个)^[1]。

由所给数据得到2个判别函数 y_1 和 y_2 。由 y_1 得到3类均值的判别分数为2 000, -1 941, -2 686, 则 $\bar{y}^{(1)} > \bar{y}^{(2)} > \bar{y}^{(3)}$, 可不重新排序, 计算得2个临界值分别为: $\bar{y}_c^{(1)} = 1 343, \bar{y}_c^{(2)} = -2 500$ 。将各样品数据代入判别函数, 得判别分数(略)。

由临界值原则可判得前1~6, 8, 10, 12样品属于第1类, 7, 9, 11, 13~18, 27样品属于第2类, 19~26样品属于第3类。错判7个(其中有6个原属于第1类错判为第2类), 判别结果的正确率为74.074%, 不到85%。易看出造成这种现象的原因是第1类有15个样品, 第2类仅有3个样品, 所以 $\bar{y}_c^{(1)}$ 明显偏于 $\bar{y}^{(1)}$ 。由此看出, 当各类样品数相差较大时, 用类间临界值来分类回判正确率不高。

由(4)式计算加权临界值为: $\bar{y}_c^{(1)} = 1 159, \bar{y}_c^{(2)} = -2 285$ 。由此临界值判别结果为1~6, 8~10, 12为第1类, 7, 11, 14~18, 27为第2类, 其余为第3类, 错判5个, 回判结果正确率为81.48%。

分别以第一、二判别函数值为横、纵坐标, 给出各样品的散点图(图1)。

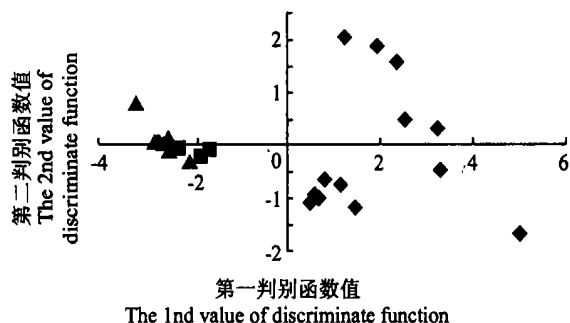


图1 陕西灌区区划样品散点图
· 类1; □ 类2; △ 类3

Fig 1 The Scatterplots of the sprinkling irrigation areas in Shaanxi
· Class 1; □ Class 2; △ Class 3

由图1可以明显看出, 只要把第1个临界值缩小, 如为0 000, 那么所有第1类样品就皆属于第1

类, 只有 27 号样品原属第 2 类被判为第 3 类, 判别正确率达 96.30%。

分别用公式(5)和(6)判别, 判别正确率分别为 92.59% (错判 2 个, 第 18 个样品判为第 3 类, 第 27 个样品判为第 2 类)和 100%, 用公式(7)判别情况和用公式(5)所得结果一致, 用公式(8)判别正确率也为 100%。

3 结论和讨论

(1)用临界值准则判别的正确率较小, 除非第一判别函数的判别效率很高, 特别当各类样品数相差悬殊时, 临界值明显偏于样品数大的一类。即使使用改进的临界值准则也未必能得到较大的改进。当总样品数不是很大, 且前两个判别函数的判别效果较好时, 用如图 1 的图示法也不难获取 1 组较好的临界

值。

(2)用马氏距离法一般可得到比临界值准则法更好的判别效果, 举例中用公式(5)计算马氏距离的判别正确率远高于用临界值准则。在用马氏距离法时, 用多个判别函数比用 1 个判别函数效果好, 即用公式(6)比用公式(5)好。按各类协方差矩阵进行分类使用合并组内协方差矩阵的判别效果好, 即用公式(6)比用公式(7)好, 因为前者充分考虑到各类协方差矩阵的区别, 但用(6)式计算量大, 一般较少用。本研究提出的新马氏距离法的公式(8)和公式(6)是等效的, 此法的主要特点是计算马氏距离用各类样品判别函数得分值而不用原始数据, 充分体现了判别分析的降维作用, 使判别分析的统计意义更明显, 有利于作进一步的统计分析。

[参考文献]

- [1] 袁志发, 孟德顺 多元统计分析[M]. 陕西杨陵: 天则出版社, 1993
- [2] 张尧庭, 方开泰 多元统计分析引论[M]. 北京: 科学出版社, 1999
- [3] 裴鑫德 多元统计分析及其应用[M]. 北京: 北京农业大学出版社, 1991

Comparison of several common discriminate criteria in Fisher discriminate analysis

ZHOU Jing-yu, SONG Shi-de, GUO Man-cai

(College of Life Sciences, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi 712100, China)

Abstract: The article discussed and compared a few common discriminate criteria used for sample categorizing in Fisher discriminate analysis. The study showed that the result of using Mahalanobis distance criteria is better than using critical value criteria in Fisher discriminate analysis. Even if using the improved critical value criteria, the result does not improve. A new Mahalanobis distance critical is shown.

Key words: Fisher discriminate; discriminate criteria; Mahalanobis distance