浙江科技学院学报，第 19 卷第 3 期，2007 年 9 月
Journal of Zhejiang University of Science and Technology
Vol. 19 No. 3, Sep. 2007

一种多总体的判别分析

胡俊娟，孙莉萍
（浙江科技学院 理学院，杭州 310023）

摘　要：判别分析是对样品或者个体进行判别分类的一种统计方法。对多总体数据进行判别分析时，总体之间的差别信息被看成同等重要，然而，这种判别效果往往不是很好。为此，提出一种结合分组的判别方法，根据各个总体的差别不同将总体分成若干组，然后对每组的数据进行判别分析，并将其应用于实例，证明在一定条件下，这种结合分组的判别方法有效。

关键词：判别分析；分组；误判概率
中图分类号：O212.4　　**文献标识码**：A　　**文章编号**：1671-8798(2007)03-0175-04

Discriminant Analysis with Multiclasses

HU Jun-juan, SUN Li-ping
(School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract: Discriminant analysis is a commonly used technique to classify a set of observations into predefined classes. The difference between classes is the same importance for us, so the effect of the discriminate analysis with multiclasses is not always good. A discrimination method is proposed by grouping under classes, which puts classes into different groups according to the different information between classes and makes discriminant analysis for groups. According to the example, the method is proved to be effective in certain condition.

Key words: discriminant analysis; grouping; misclassified probability

判别分析是多元统计分析中实用分析方法之一。随着电子技术的发展，人们常常需要对获得的资料数据进行判别分析，所以判别分析方法很受人们的关注，在实际生活中有着广泛的应用[1,2]。目前，对于判别分析的讨论主要是关于样品数据[3-5]或者判别方法等方面[6-8]。实用的判别分析方法主要是距离判别和 Bayes 判别、Fisher 判别[9]。对于各种判别方法而言，总体间的差别往往被看作等同。在总体间差别不同的时候，这种一般判别方法的效果可能就不太好。分组分层技术是一项非常有效的技术，通过分组分层及与其他方法的结合，往往可以使结果更加可靠[10]。因此，对于多个总体所给出的信息不同时，本文给出了结合分组的方法进行判别。

收稿日期：2007-03-09

作者简介：胡俊娟(1979—　)，女，浙江兰溪人，讲师，浙江大学在读硕士研究生，主要从事基础数学和数理统计的教学和研究。

1 原 理

对多总体的数据进行判别分析，分组可以很好地利用各变量在数据中的不同影响，对样品的判别起到一定的改进作用。例如：对一件衣服的尺码分类（如：加大、大、中、小），有袖长、身长、胸围、领围等指标，根据经验知道身长、胸围是重要指标，可以根据这2个指标把加大号和大号结合，将中号和小号尺码结合分成2组，再结合各个变量对这2组进行判别分类，然后将原来结合的各个类分开再进行判别。为了使判别分析的效果更好，可根据对于类可分离的变量进行分组，以便更好地说明不同变量在判别中的不同作用。因此，采用可分离变量进行分组的方法：先将各个变量与类作图，从图中直观上得出将类分离的变量（可分离变量），用这些变量将类分组，把聚集的类合成一组，对不同的组进行方差分析，以得出对于不同的组变量是否有显著性差异。这样就可以找到一种使组内方差最小的分组方法，组与组之间的判别分析就能达到较好的效果。

假定 $G_1, G_2, \cdots, G_g$ 均为 p 维正态总体，在等误判损失和协方差矩阵相等的条件下，Bayes 判别准则简化为：

$$x \in G_i \text{ 若 } \max_{1 \leqslant k \leqslant g} \{d_k(x)\} = d_i(x),$$

其中 $d_i(x) = \mu_i^{\mathrm{T}} \Sigma^{-1} x - \frac{1}{2} \mu_i^{\mathrm{T}} \Sigma^{-1} \mu_i + \ln q_i$，

μ_i, Σ_i, q_i 分别为 G_i 的期望、协方差和先验概率。实际应用中，若 μ_i, Σ_i 未知，则可用训练样本的样本均值 $\bar{x}^{(i)}$ 和样本方差 S_i 作为 μ_i 和 Σ_i 的估计，用 S_i 联合估计 Σ，即[11]

$$\hat{\Sigma} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{n_1 + n_2 + \cdots + n_g - g}$$

从判别准则可以得出，对于同一组的类，从判别函数可以得出类之间的距离比原来的距离大，有利于组内各个类的判别。如果将多总体通过总体间的差别程度进行分组，组间距离可以很好地将各组分开。

在实际应用中，可利用的资料只是来自各总体的训练样本，而总体的分布是未知的，通常评价准则是以训练样本为基础的貌似误判率方法和刀切法（交叉确认法），利用回判的误判率来衡量判别准则的效果，验证判别分析的结果。

2 应 用

考虑下面的判别分析：某商学院在招收研究生时，以学生在大学期间的平均学分（GPA）和管理能力考试成绩（GMAT）帮助录取研究生，对申请者划归为三类，G1：录取，G2：未录取，G3：待定[11]。

由于事先对于各个变量的特征和分类方法没有任何的专业知识，所以，首先尝试对数据进行直接的判别分类，利用 Bayes 判别准则对训练样本进行回判[12]。在判定损失相等、总体先验分布相同且三个总体服从协方差矩阵相等的正态分布的假定下建立 Bayes 判别准则。由 SAS 程序可输出部分结果如下：

The DISCRIM Procedure

Classified

Obs	From category	into category		G1	G2	G3
2	G1	G3	*	0.1201	0.0020	0.8779
3	G1	G3	*	0.3652	0.0004	0.6344
24	G1	G3	*	0.4769	0.0000	0.5230
31	G1	G3	*	0.2963	0.0004	0.7033
58	G2	G3	*	0.0001	0.2453	0.7546
59	G2	G3	*	0.0001	0.1328	0.8671
66	G3	G1	*	0.5343	0.0000	0.4657

* Misclassified observation

Number of Observations and

Percent Classified into category

From category	G1	G2	G3	Total
G1	27	0	4	31
	87.10	0.00	12.90	100.00
G2	0	26	2	28
	0.00	92.86	7.14	100.00
G3	1	0	25	26
	3.85	0.00	96.15	100.00

因此在等误判损失和协方差矩阵相等的条件下，用 Bayes 判别准则得到其貌似误判率为：7/85＝0.082。若利用刀切法评估此准则的优良性，用 SAS 程序可得到其误判率为：9/85＝0.106。

针对多总体判别分析，结合分组的判别法首先尝试对数据作直观图，分别作 GPA、GMAT 和类之柱形图图1、图2。

根据柱形图图1，对变量 GPA 来说，G1 与 G2 并成一组，G3 为一组。对这2个分组作双因子（一因子为2个变量，另一因子为组）方差分析，部分结果如下：

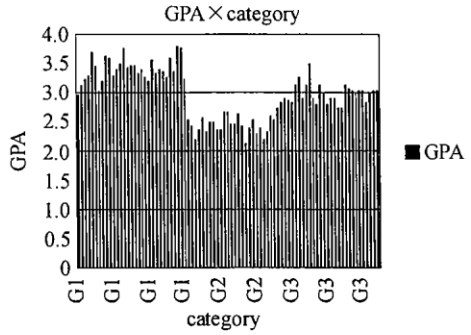

图 1　类(总体)与 GPA 的柱形图

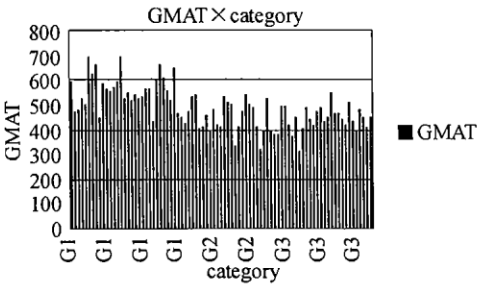

图 2　类(总体)与 GMAT 的柱形图

The ANOVA Procedure

Source	DF	Anova SS	Mean Square	F Value	Pr>F
a	1	10017520.56	10017520.56	3416.48	<.0001
B	1	36636.90	36636.90	12.50	0.0005
a*B	1	34935.65	34935.65	11.91	0.0007

对变量 GMAT 而言，从图 2 中看出可以把 G2 与 G3 并成一组，G1 为一组。对这 2 个组作方差分析，得到其结果为：

Source	DF	Anova SS	Mean Square	F Value	Pr > F
a	1	10017520.56	10017520.56	5543.01	<.0001
B	1	130675.78	130675.78	72.31	<.0001
a*B	1	127627.71	127627.71	70.62	<.0001

比较上述 2 个方差分析中显著性分析(Pr>F 的值)得出用第二种分法更好，所以把 G1，G2，G3 分成 G1 和 G2G3 两层，对这两层进行判别分析。用 SAS 输出部分结果为：

Classified

Obs	From A	into A		1	2
2	1	2	*	0.3563	0.6437
66	2	1	*	0.7188	0.2812

Number of Observations and Percent Classified into A

From A	1	2	Total
1	30	1	31
	96.77	3.23	100.00
2	1	53	54
	1.85	98.15	100.00
Total	31	54	85
	36.47	63.53	100.00

从结果可以看出，G1 中第 2 号被误判入待定的 G2G3 结合的类，而 G3 中的 66 号被误判入 G1 中，从而貌似误判率为 2/85＝0.023 5，同样可用 SAS 输出刀切法的误判概率也为 2/85＝0.023 5。接下来对 G2G3 结合的这一层进行判别，有结果如下：

Classified

Obs	From category	into category		G2	G3
27	G2	G3	*	0.3080	0.6920
28	G2	G3	*	0.1466	0.8534

* Misclassified observation

Number of Observations and Percent Classified into category

From

category	G2	G3	Total
G2	26	2	28
	92.86	7.14	100.00
G3	0	26	26
	0.00	100.00	100.00
Total	26	28	54
	48.15	51.85	100.00

上述结果给出了按判别准则给出的误判概率为：2/85＝0.023 5。若利用刀切法评估此准则的优良性，计算其误判率结果还是为 2/85＝0.023 5。

由上述的这两步可以知道：结合分组的判别分析法可以有效地减小误判概率，使判别分析结果更有效。给出一个新的待判别的样品，先判别它是属于第一步中的第一组(G1)，还是第二组(G2G3)，如果样品属于第二组，再判别其属于第二组中的哪个总体。

3　结　语

从实例可以看出，用直接判别方法误判概率不小于 0.082，而用结合分组的判别分析方法误判概率不大于 2/85＋2/85＝0.047。因而对新的样品(个体)进行判别归类，用结合分组的判别方法可以有效地减小误判概率，可见，这种判别分析方法能显著提高判别质量。判别有效的前提是对于不同的类中哪几个类可以看成一组，本文的方差分析只是提供了一种方法来解决这个问题，还可以结合类之间的距离来考虑，距离相对较近的考虑分成一组。具体问题要进行具体分析，以找到最合适的分组类型。

(下转第 196 页)

(上接第177页)

参考文献:

[1] 乔辉辉,牛芳.上市公司财务困境预测的Fisher判别分析模型[J].统计与信息论坛,2003,18(2):69-71.

[2] 施锡铨,邹新月.典型判别分析在企业信用风险评估中的应用[J].财经研究,2001,27(10):53-57.

[3] TRENDAFILOV NICKOLAY T, JOLLIFFE IAN T. DALASS: Variable selection in discriminant analysis [J]. Computational Statistics & Data Analysis, 2007, 51:3718-3736.

[4] HUANG Yu-fen, KAO Tzu-Ling, WANG Tai-Ho. Influence functions and local influence in linear discriminate analysis[J]. Computational Statistics & Data Analysis, 2007,51:3844-3861.

[5] POON Wai-yin. Identifying influential observations in logistic discriminant analysis[J]. Statistics & Probability Letters, 2006, 76:1348-1355.

[6] DAI Dao-qing, YUEN P C. Wavelet based discriminant analysis for face recognition[J]. Applied Mathematics and Computation, 2006, 175:307-318.

[7] BOUMAZA R. Discriminant analysis with indepently repeated multivariate measurements: an L^2 approach [J]. Computational Statistics & Data Analysis, 2004, 47:823-843.

[8] MIA Hubert, DRIESSEN KATRIEN Van. Fast and robust discriminant analysis[J]. Computational Statistics & Data Analysis, 2004, 45:301-320.

[9] 何晓群.多元统计分析[M].北京:中国人民大学出版社,2004:100-105.

[10] 李培军.不等概率抽样估计的应用[J].辽宁师范大学学报:自然科学版,2004,27(4):385-388.

[11] 梅长林,周家良.实用统计方法[M].北京:科学出版社,2002:108.

[12] JOHNSON R A, WICHERN D W. Applied Multivariate Statistical Analysis [M]. 3rd ed. Englewood cliffs, New Jersey: Prentice-Hall Inc, 1992:535-536.