文章编号: 1002–1566 (2014) 02 – 0256 – 10

判别分析与 Logistic 回归组合分类

尹剑 陆程敏 杨贵军

(天津财经大学中国经济统计研究中心, 天津 300222)

摘要: 本文讨论判别分析与 Logistic 回归组合方法的二分类问题。模拟结果显示, 基于判别分析的 Logistic 回归的第一组合法的回判正确率往往高于判别分析分类和 Logistic 回归分类。第二组合法是基于 Logistic 回归的判别分析分类, 回判正确率接近于 Logistic 回归的分类结果。很多情况下, 这两种方法的回判正确率都高于判别分析。

关键词: 判别分析; Logistic 回归; 回判正确率

中图分类号: F224.7, O212.1 **文献标识码**: A

http://www.ixueshu.com

Combinations of Discriminatory Analysis and Logistic Regression for Classification

YIN Jian LU Cheng-min YANG Gui-jun

(Tianjin University of Finance and Economics, China Center for Economics Statistics Research, Tianjin 300222, China)

Abstract: The combinations of discriminatory analysis and logistic regression could improve to solve binary classification problems. Simulation results show that discriminatory accuracy rate of logistic regression with classification of discriminatory analysis based often more large than ones of discriminatory analysis or logistic regression. Discriminatory accuracy rate of discriminatory analysis with classification of logistic regression approximately equals to one of logistic regression. In many situations, the two methods are often better than discriminatory analysis.

Key words: discriminatory analysis, logistic regression, discriminatory accuracy rate

0 引言

分类问题在社会科学和自然科学的研究中经常遇到。比如, 银行对信用卡客户进行信用等级分类, 通常会考虑诸如年龄、职业、收入、房产、婚否等指标。上市公司股票类型的划分也往往会依据该公司财务指标。最常用的分类方法是判别分析分类和 Logistic 回归分类。判别分析分类依据观测值与不同类别间距离差异, 而 Logistic 回归分类依据 Logistic 回归的拟合值。在医学、生物学和经济管理等诸多领域, 判别分析和 Logistic 回归都有着广泛的应用。

收稿日期: 2011 年 12 月 23 日 **收到修改稿日期**: 2012 年 4 月 17 日

基金项目: 教育部新世纪优秀人才支持计划 (NCET-08-0909), 教育部留学回国人员科研启动基金项目, 全国统计科学研究计划项目 (2010LC60、2011LY096), 天津市哲学社会科学规划项目 (TJTJ12-004), 天津财经大学科研发展基金项目 (Y1204)。

Biometrics、Biometrical Journal 等杂志刊登了很多判别分析和 Logistic 回归的生物医学论文。范书山等 [1] 应用 Logistic 回归对全胃肠外营养中心静脉导管感染危险因素进行了研究。朱燕波等 [2] 对中医体质类型与超重/肥胖关系的进行了 Logistic 回归分析。郭万越 [3] 应用 Logistic 回归分析了非酒精性脂肪肝相关因素及预防政策。田恒宇 [4] 应用判别分析对亚健康状态常见证候特征进行研究。Zavgren[5] 选用 Logistic 回归和判别分析对保险公司的破产原因进行实证分析。Lee、Hyun and Urrutia[6] 利用 Logistic 回归对非寿险公司偿付能力进行了研究。王春峰等 [7] 将判别分析和 Logistic 回归应用于我国商业银行信用风险评估，并对两种方法进行了比较。梁琪 [8] 应用 logistic 回归对企业经营进行预警。龚承刚 [9] 应用 Logistic 回归评价企业竞争力。缪瑾等 [10] 用 Logistic 回归模型分析了性别和不同年级的大学生作弊的主要因素。雷庆祝、刘诗茂 [11] 利用 Logistic 回归分析方法对传统粉笔字教学和多媒体教学方式的选择建立模型并对影响教学的各因素进行了分析。陈磊、任若恩 [12] 用判别分析估计财务比率变化，研究公司财务危险预警。张阔 [13] 等运用判别分析和 Logistic 回归分析消费者寿险购买行为，同时指出将 Logistic 回归模型和判别模型结合起来的联合预测模型的预测精度更高。

判别分析和 Logistic 回归为实际分类问题提供有价值的分类信息。相比较而言，Logistic 回归更稳健，但并没有理论结果显示该方法回判正确率更高。回判正确率指正确分类观测值的所占比例。张初兵、高康和杨贵军 [14] 认为，判别分析的回判正确率并不总是与 Logistic 回归的回判正确率一致，甚至有时差异很大。在很多情况下，Logistic 回归比判别分析稳健，回判正确率也更高。

本文针对二分类问题，研究 Logistic 回归和判别分析组合方法的回判正确率。本文考虑两种组合形式，一种是先对数据进行判别分析，再利用判别分析的结果进行 Logistic 回归分析；另一种是先对数据进行 Logistic 回归，再利用 Logistic 回归结果进行判别分析。随机模拟的结果显示，基于判别分析的 Logistic 回归的回判正确率往往高于只进行判别分析或 Logistic 回归的分类结果。基于 Logistic 回归的判别分析的回判正确率接近于 Logistic 回归的分类结果。很多情况下，这两种方法的回判正确率都高于判别分析。

1 判别分析和 Logistic 回归的组合方法

二分类问题考虑来自两总体的样本。设两个互不相交的总体分别为 R_1 和 R_2，$R_1 \cup R_2 = \Omega$，$R_1 \cap R_2 = \emptyset$，Ω 为所有样本点集合，均值向量为 μ_1 和 μ_2，协方差阵为 Σ_1 和 Σ_2。

本文的判别分析是基于马氏距离，将观测点判定为来自距离总体均值“最短”的总体。令 $x \in R_1$，则 x 与总体 R_1 的马氏距离为 $d(x, R_1) = \sqrt{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}$。当 $\Sigma_1 = \Sigma_2 = \Sigma$ 时，取 $\omega(x) = d^2(x, R_2) - d^2(x, R_1) = 2(x-\overline{\mu})^T\Sigma^{-1}(\mu_1 - \mu_2)$ 为判别函数；当 $\Sigma_1 \neq \Sigma_2$ 时，$\omega(x) = (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)$，其中 $\overline{\mu} = (\mu_1+\mu_2)/2$。判别规则为

$$\begin{cases} x \in R_1, & \text{如果 } \omega(x) \geq 0, \\ x \in R_2, & \text{如果 } \omega(x) < 0. \end{cases} \tag{1}$$

当总体均值和协方差阵未知，用样本均值和样本协方差阵作为其估计值。关于判别分析的更多内容请参见张尧庭和方开泰 [15]，王学仁、王松桂 [16]，张润楚 [17] 等。

常用二分类的 Logistic 回归模型的响应变量 Y 取值为 0 和 1，模型为

$$\ln\big(\Pr(Y=1)/(1-\Pr(Y=1))\big) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

其中 $X_1,\cdots,X_p$ 为解释变量，ε 为误差项。本文约定 ε 服从伯努利分布。Logistic 回归的参数主要用极大似然估计法，记其为 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$。对于 $x=(x_1,\cdots,x_p)\in\Omega$，概率 $\Pr(y=1|x)=e^{\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p}/\left(1+e^{\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p}\right)$。分类规则为

$$\begin{cases} x\in R_1, & \text{如果 } \Pr(y=1|x)\geq 0.5,\\ x\in R_2, & \text{如果 } \Pr(y=1|x)<0.5. \end{cases} \tag{2}$$

有关 Logistic 回归的内容可参见 Agrest[18]。

综合上述讨论，判别分析是依据观测样本 x 的判别函数 $\omega(x)$ 进行分类。$\omega(x)$ 的绝对值越大，误判的可能性越低；$\omega(x)$ 的绝对值较小，误判的可能性越高。而 Logistic 回归是根据条件概率 $\Pr(y=1|x)$ 与 0.5 差的绝对值对观测数据进行分类，差的绝对值越大，误判的可能性越低；差的绝对值越小，误判的可能性越高。张初兵、高康和杨贵军 (2010) 指出，判别函数 $\omega(x)$ 的绝对值越大，往往 $\Pr(y=1|x)$ 与 0.5 差的绝对值也越大；反之，$\Pr(y=1|x)$ 与 0.5 差的绝对值越大，判别函数 $\omega(x)$ 的绝对值也越大。但是，两种方法分类结果并不总是一致的。可以利用判别分析结果，改进 Logistic 回归的分类，也可以利用 Logistic 回归结果改进判别分析的分类。下面介绍判别分析和 Logistic 回归组合分类方法。

1.1 基于判别分析的 Logistic 回归分类方法

对于样本数据，先进行判别分析，再利用判别分析结果对 Logistic 回归改进的分类方法，简称为第一组合法。

基于正态分布数据分类的模拟经验，当 $\omega(x)$ 的函数值大于 85% 分位数和小于 15% 分位数时，判别分析与 Logistic 回归的分类结果几乎一致。在很多情况下，分类结果都正确，不需要修正。在下文模拟中，将样本数据的判别函数 $\omega(x)$ 值由小到大依次排列，定义函数 $d(x)$

$$d(x)=\begin{cases} 0, & \text{如果 } \omega(x) \text{ 大于 85\% 分位数或小于 15\% 分位数},\\ \omega(x), & \text{如果 } \omega(x) \text{ 在 15\% 分位数和 85\% 分位数之间}. \end{cases} \tag{3}$$

当 $\omega(x)$ 值大于 85% 分位数或小于 15% 分位数时，$d(x)$ 值设定为 0，还可以减少 $\omega(x)$ 极大值和极小值对 Logistic 回归的影响。将 $d(x)$ 样本值作为 Logistic 回归的一个新增解释变量，建立 Logistic 回归模型，对观测数据进行分类。

1.2 基于 Logistic 回归的判别分析方法

第二组合法是先建立 Logistic 回归模型，再用 Logistic 回归结果改进判别分析的分类。

与第一组合法相似，基于正态分布数据分类的模拟经验，当 $\Pr(y=1|x)$ 值大于 85% 分位数和小于 15% 分位数时，Logistic 回归的分类结果往往是正确的。在下文模拟中，将 Logistic 回归的拟合值由小到大排列，定义函数 $l(x)$

$$l(x)=\begin{cases} \Pr(y=1|x), & \text{如果 } \Pr(y=1|x) \text{ 位于 15\% 分位数与 85\% 分位数之间},\\ 0, & \text{如果 } \Pr(y=1|x) \text{ 小于 15\% 分位数或大于 85\% 分位数}, \end{cases} \tag{4}$$

将 $l(x)$ 的样本值作为判别分析的一个新增解释变量，利用判别分析对观测数据进行分类。

本文的函数 $d(x)$ 和 $l(x)$ 选择更多依赖于正态分布数据的模拟经验。在实际应用中，函数 $d(x)$ 和 $l(x)$ 的设定还需要考虑观测数据的分布类型等因素。

http://www.ykxshu.com

2 判别分析与 Logistic 回归的组合分类法的模拟比较

为了评价第一组合法和第二组合法，本节选用判别分析和 Logistic 回归进行对比。采用随机模拟的方法，对相同样本数据应用四种方法进行分类，计算回判正确率。更优分类方法的回判正确率更高。

在实际应用中，很多观测数据都用正态分布近似。有时，尽管观测数据不是来源于正态分布，也可以进行正态化变化，变换为服从或近似服从正态分布的观测数据。下文的观测数据都利用标准正态分布产生的随机数。变量间的相关性会影响数据分类结果，共线性存在会影响 Logistic 回归方法的应用。这里，只考虑变量之间相互独立的情况。数据类别是观测变量的线性函数，也可能是非线性函数。实际应用中，往往选择线性函数，即使不是线性函数，也常采用线性函数近似。在下文模拟中，选择数据类别是观测变量的线性函数。

2.1 样本数据的生成

定义线性函数

$$f(X_1,\cdot,X_p)=b_0+b_1X_1+\cdots+b_pX_p+\varepsilon,$$

其中 $b_0,b_1,\cdots,b_p$ 为系数，用标准正态分布随机数代替，$\varepsilon\sim N(0,\sigma^2)$。$X_1,\cdots,X_p$ 为观测变量，随机样本生成过程如下：

(1) 生成 $p+1$ 个随机数，依次赋值给 $b_0,b_1,\cdots,b_p$，随机数服从标准正态分布；

(2) 生成 n 组 p 维标准正态分布的随机数，依次赋值给 $x_{i1},\cdots,x_{ip}$，$i=1,\cdots,n$；

(3) 生成 n 个服从均值为 0 且方差为 σ^2 的正态分布随机数，赋值给 $\varepsilon_1,\cdots,\varepsilon_n$；

(4) 计算函数值 $f_i(x_{i1},\cdots,x_{ip})=b_0+b_1x_{i1}+\cdots+b_px_{ip}+\varepsilon_i$，并计算函数值的中位数，记为 $Median(f)$；

(5) 对于 $i=1,\cdots,n$，令

$$y_i=\begin{cases}1, & \text{如果 } f_i(x_{i1},\cdots,x_{ip})\geq Median(f),\\ 0, & \text{如果 } f_i(x_{i1},\cdots,x_{ip})<Median(f),\end{cases}$$

则 $y_i,x_{i1},\cdots,x_{ip}$，$i=1,\cdots,n$，即为所分析的样本数据。$y_i=1$ 与类别 1 对应；$y_i=0$ 类别 0 保证。保证两个类别的样本容量相等。

按照上述过程产生样本数据，分别利用四种方法进行分类，计算每种方法的回判正确率。分析参数 p,n,σ^2 的不同取值对四种方法的回判正确率的影响。

2.2 组合分类法的模拟结果

为了演示分类方法的优劣，这里给出了每种方法回判正确率均值和最优频数。回判正确率均值是利用 200 组模拟数据，某分类方法的回判正确率的平均数；最优频数是对于 200 组模拟数据，该分类方法的回判正确率为最大值的频数。模拟研究经验显示，当模拟次数超过 200 时，随着模拟次数增加，四种方法的回判正确率并没有显著变化，下文只给出了参数变化时每种方法的 200 次模拟结果。由于解释变量个数 p、方差 σ^2 和样本量 n 对分类结果影响大，这里选择的 n 分别为 50,100,150,200，σ^2 分别为 1,0.25,0.04,0.01，p 分别为 2,3,5,10。

(1) 解释变量个数 $p=2$ 的模拟结果

解释变量个数 $p=2$ 的模拟结果见表 1。表 1 第 2 列依次给出不同 σ^2 时样本量 n 分别为 50,100,150,200 的回判正确率，第 3 列依次对应给出了最优次数。每个括号内的四个数值依次

http://www.ixueshu.com

对应判别分析、Logistic 分类、第一组合方法和第二组合方法的模拟结果。

表 1 $p=2$ 时四种方法的回判正确率和最优频数

σ^2	回判正确率	最优频数
1	(0.7672, 0.7668, 0.7675, 0.7679)	(98 92 89 94)
	(0.7584, 0.7593, 0.7593, 0.7580)	(82 82 84 77)
	(0.7520, 0.7518, 0.7519, 0.7538)	(66 64 70 85)
	(0.7627, 0.7628, 0.7626, 0.7623)	(80 67 73 77)
0.25	(0.8493, 0.8522, 0.8524, 0.8512)	(87 94 107 101)
	(0.8579, 0.8588, 0.8588, 0.8588)	(88 86 93 101)
	(0.8547, 0.8550, 0.8553, 0.8550)	(82 73 82 77)
	(0.8382, 0.8380, 0.8376, 0.8374)	(71 87 80 71)
0.04	(0.9235, 0.9304, 0.9316, 0.9299)	(72 111 135 93)
	(0.9335, 0.9357, 0.9372, 0.9346)	(72 108 113 94)
	(0.9242, 0.9267, 0.9267, 0.9264)	(60 76 92 75)
	(0.9341, 0.9365, 0.9368, 0.9352)	(62 93 104 78)
0.01	(0.9541, 0.9648, 0.9661, 0.9632)	(49 135 155 126)
	(0.9564, 0.9629, 0.9635, 0.9626)	(59 115 121 100)
	(0.9617, 0.9655, 0.9659, 0.9647)	(46 111 132 94)
	(0.9603, 0.9642, 0.9647, 0.9637)	(43 107 119 76)

表 1 显示，当 $\sigma^2=1$ 且 $n=50$，四种方法的回判正确率为 0.7672、0.7668、0.7675 和 0.7679，方法之间的差异小，第二组合法的回判正确率略高。当 $\sigma^2=1$ 和 $n=100$，Logistic 分类和第一组合法的回判正确率略高。当 $\sigma^2=1$ 和 $n=150$，第二组合法的回判正确率略高。当 $\sigma^2=1$ 和 $n=200$，Logistic 分类的回判正确率最高，略优于第一组合法或第二组合法的回判正确率。当 $\sigma^2=1$ 且 $n=50$，四种方法的回判正确率为最大值的频数分别为 98、92、89 和 94，方法之间的差异小，判别分析的回判正确率为最大值的频数最大，不到模拟次数 50%。第二组合法的回判正确率为最大的频数次之。当 $\sigma^2=1$ 和 $n=100$，第一组合法的回判正确率为最大值的频数最大。当 $\sigma^2=1$ 和 $n=150$，第二组合法的回判正确率为最大值的频数最大。当 $\sigma^2=1$ 和 $n=200$，判别分析的回判正确率达到最大值的频数最大。当 $\sigma^2=1$ 时，四种方法的回判正确率没有明显差异，第一组合法和第二组合法的分类结果有时更好。

随着方差 σ^2 减小，四种方法的回判正确率都明显增大。当 $\sigma^2=0.25$ 和 $n=50$、150，第一组合法的回判正确率均值分别增长到 0.8524、0.8553，回判正确率为最大值的频数分别为 107、82，是两种情况下回判正确率均值和最优频数的最大值。当 $\sigma^2=0.25$ 和 $n=100$，Logistic 回归、第一组合法和第二组合法的回判正确率均值相同，为 0.8588，第二组合法的最优频数最大，为 101。当 $\sigma^2=0.25$ 和 $n=200$，判别分析的回判正确率均值略高，为 0.8382，Logistic 分类的最优频数最大，为 87。当 $\sigma^2=0.04$，第一组合法的回判正确率均值分别增长到 0.9316、0.9372、0.9267 和 0.9368，回判正确率的最优频数分别为 135、113、92 和 104，高于其它 3 种方法的回判正确率均值和最优频数。第二组合法的回判正确率均值和最优频数接近于 Logistic 回归，判别分析的回判正确率均值和最优频数最低。当 $\sigma^2=0.01$，第一组合法的回判正确率均值和最优频数都为最大，第二组合法的回判正确率均值和最优频数接近于 Logistic 分类，判别分析的回判正确率均值和最优频数最低。

图 1 给出了 $n=200$ 不同方差的 4 种分类方法的回判正确率盒子图。方差 σ^2 的不同值在横坐标用数值标出，纵坐标为回判正确率。每组盒子图从左到右依次为判别分析、Logistic 回

归、第一组合法和第二组合法。图 1 显示，随着 σ^2 增大，4 种分类法的回判正确率都减小，回判正确率波动性增大。相比较而言，第一组合法的盒子高度较小，回判正确率波动性较小。

图 2 给出了 $\sigma^2 = 0.04$ 不同样本量的四种分类方法的回判正确率盒子图。每类含样本点的个数，也就是样本量 n 的一半，在横坐标用数值表示，纵坐标为回判正确率。每组盒子图从左到右依次代表判别分析、Logistic 回归、第一组合法和第二组合法。图 2 显示，随着样本量变化，回判正确率变化不大，第一组合法的回判正确率均值略优于其它方法，波动性稍大。

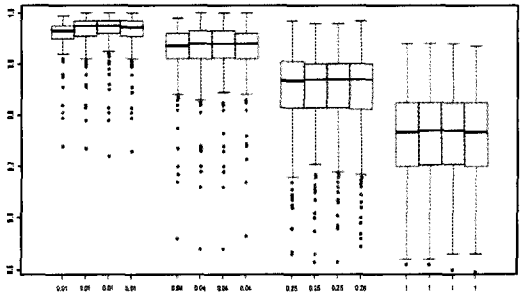

图 1 $p = 2$ 时方差不同时四种分类方法回判正确率的盒子图

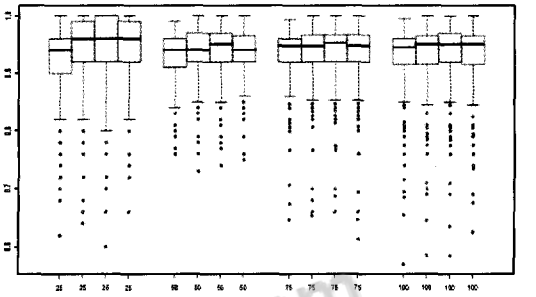

图 2 $p = 2$ 时样本量不同时四种分类方法回判正确率的盒子图

(2) 解释变量个数 $p = 3$ 的模拟结果

表 2 $p = 3$ 时四种方法的回判正确率和最优频数

σ^2	回判正确率	最优频数
1	(0.8249, 0.8176, 0.8264, 0.8313)	(94 84 97 110)
	(0.8161, 0.8131, 0.8151, 0.8165)	(90 73 83 80)
	(0.8069, 0.8071, 0.8073, 0.8079)	(92 64 70 85)
	(0.8086, 0.8072, 0.8076, 0.8096)	(85 63 73 77)
0.25	(0.8934, 0.9004, 0.9060, 0.9032)	(84 108 132 101)
	(0.8897, 0.8931, 0.8959, 0.8946)	(72 91 108 85)
	(0.8898, 0.8903, 0.8914, 0.8904)	(80 75 97 82)
	(0.8875, 0.8899, 0.8911, 0.8901)	(70 74 80 75)
0.04	(0.9447, 0.9660, 0.9706, 0.9650)	(41 146 186 145)
	(0.9458, 0.9567, 0.9601, 0.9565)	(43 118 156 116)
	(0.9452, 0.9542, 0.9562, 0.9537)	(44 92 130 81)
	(0.9435, 0.9497, 0.9509, 0.9486)	(42 85 123 79)
0.01	(0.9560, 0.9880, 0.9917, 0.9885)	(30 178 191 171)
	(0.9618, 0.9837, 0.9867, 0.9827)	(22 140 172 132)
	(0.9627, 0.9779, 0.9797, 0.9766)	(23 143 159 126)
	(0.9641, 0.9778, 0.9799, 0.9774)	(24 114 162 94)

解释变量个数 $p = 3$ 的模拟结果见表 2，表 2 的结构同表 1。表 2 显示，当 $\sigma^2 = 1$，第二组合法的回判正确率均值最高，分别为 0.8313、0.8165、0.8079 和 0.8096，四种方法的回判正确率差异较小。当 $\sigma^2 = 1$ 和 $n = 50$，第二组合法的最优频数最大，为 110，而其他情况下，判别分析的最优频数最大，分别为 90，92，85，小于模拟次数的 50%。

随着方差 σ^2 减小，四种方法的回判正确率都增大，增大的幅度小于 $p = 2$ 的情况。当 $\sigma^2 = 0.25$、0.04 和 0.01，第一组合法的回判正确率均值最大，最优频数也最大。当 $\sigma^2 = 0.04$ 和 $n = 50$ 与当 $\sigma^2 = 0.01$ 和 $n = 50$，第一组合法的最优频数分别为 186 和 191，超过了模拟次

http://www.ixueshu.com

数的 90%。第二组合法的回判正确率均值和最优频数接近于 Logistic 回归，判别分析的回判正确率均值和最优频数都相对最小。

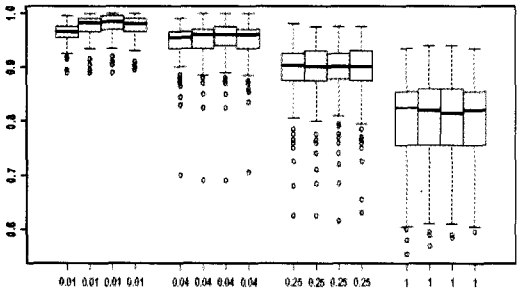

图 3 $p=3$ 时方差不同时四种分类方法回判正确率的盒子图

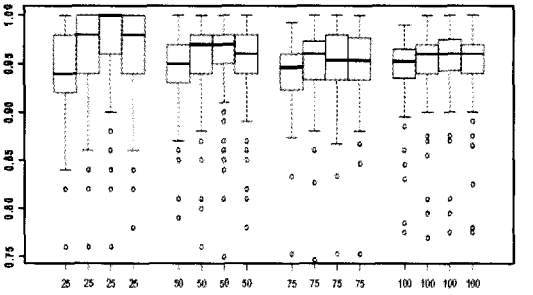

图 4 $p=3$ 时样本量不同时四种分类方法回判正确率的盒子图

图 3 给出了 $n=200$ 不同方差的 4 种方法的回判正确率盒子图。图 3 的坐标轴、盒子图与分类方法的对应关系同图 1。图 3 显示，随着 σ^2 增大，4 种分类法的回判正确率都减小。当 $\sigma^2=1$，4 种方法的回判正确率中位数和波动性的差异都比较小。当 $\sigma^2=0.25$，4 种方法的回判正确率中位数差异小，第一组合法回判正确率的波动性小，第二组合法的回判正确率波动性接近于 Logistic 回归。当 $\sigma^2=0.04$、0.01，判别分析的中位数相对较小，回判正确率波动性小，第一组合法的回判正确率中位数较大，波动性大。

图 4 给出了 $\sigma^2=0.04$ 不同样本量的四种分类方法的回判正确率盒子图。图 4 的坐标轴、盒子图与分类方法的对应关系同图 2。图 4 显示，随着样本量变化，回判正确率波动性变化小，判别分析的回判正确率中位数小。当 $n=50$、100，第一组合法的回判正确率中位数略高，波动性更小。Logistic 回归和第二组合法的回判正确率中位数接近，波动性也接近。

(3) 解释变量个数 $p=5$ 的模拟结果

表 3 $p=5$ 时四种方法的回判正确率和最优频数

σ^2	回判正确率	最优频数
1	(0.8826, 0.8720, 0.8891, 0.8903)	(66 68 111 105)
	(0.8644, 0.8595, 0.8669, 0.8683)	(74 49 82 100)
	(0.8596, 0.8551, 0.8600, 0.8619)	(79 48 70 89)
	(0.8561, 0.8525, 0.8549, 0.8564)	(87 52 73 89)
0.25	(0.9329, 0.9479, 0.9597, 0.9537)	(40 118 158 134)
	(0.9198, 0.9290, 0.9369, 0.9307)	(43 76 140 81)
	(0.9214, 0.9251, 0.9301, 0.9267)	(55 72 110 75)
	(0.9198, 0.9254, 0.9275, 0.9254)	(47 84 107 82)
0.04	(0.9546, 0.9895, 0.9930, 0.9898)	(18 182 194 180)
	(0.9532, 0.9794, 0.9838, 0.9786)	(8 138 184 131)
	(0.9547, 0.9720, 0.9762, 0.9721)	(19 104 156 104)
	(0.9571, 0.9720, 0.9752, 0.9722)	(17 100 149 93)
0.01	(0.9602, 0.9977, 0.9984, 0.9981)	(25 194 197 194)
	(0.9601, 0.9946, 0.9963, 0.9941)	(5 178 197 176)
	(0.9629, 0.9932, 0.9952, 0.9927)	(1 159 186 149)
	(0.9642, 0.9905, 0.9920, 0.9899)	(2 157 185 146)

表 3 给出了解释变量个数 $p=5$ 的模拟结果，结构同表 1。表 3 显示，当 $\sigma^2=1$，四种方

法的回判正确率差异较小，第二组合法的回判正确率均值最高，分别为 0.8903、0.8683、0.8619 和 0.8564，其中当 $n=50$，第一组合法的最优频数最大，为 111，第二组合法的最优频数次大，为 105；当 $n=100$、150 和 200，第二组合法的最优频数最大，分别为 100、89 和 89。

随着 σ^2 减小，四种方法的回判正确率都增大，增大的幅度小于 $p=3$ 的情况。当 $\sigma^2=0.25$、0.04 和 0.01，第一组合法的回判正确率均值最大，最优频数也最大。当 $\sigma^2=0.04$ 且 $n=50$、100 和当 $\sigma^2=0.01$，第一组合法的最优频数超过了模拟次数的 90%。第二组合法的回判正确率均值和最优频数接近于 Logistic 分类，判别分析的回判正确率均值和最优频数都相对最小。

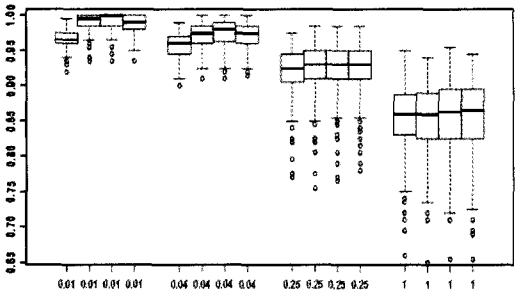

图 5 $p=5$ 时方差不同时四种分类方法回判正确率的盒子图

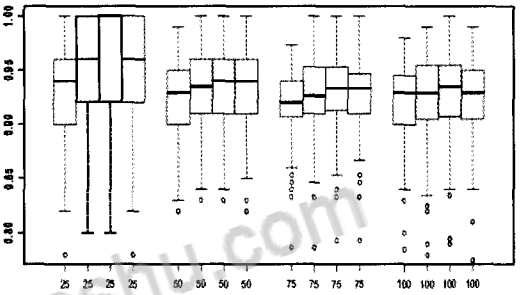

图 6 $p=5$ 时样本量不同时四种分类方法回判正确率的盒子图

图 5 给出了样本量为 200 不同方差的 4 种方法的回判正确率盒子图。图 5 的坐标轴、盒子图与分类方法的对应关系同图 1。图 5 显示，随着 σ^2 增大，4 种分类法的回判正确率都减小。当 $\sigma^2=1$，4 种方法的回判正确率中位数的差异小，第一组合法和第二组合法的回判正确率波动性大，判别分析的回判正确率波动性小。当 $\sigma^2=0.25$，4 种方法的回判正确率中位数差异小，波动性也较为接近。当 $\sigma^2=0.04$、0.01，4 种方法的回判正确率波动性接近，第一组合法的回判正确率中位数较大，第二组合法的回判正确率波动性接近于 Logistic 回归。

图 6 给出了方差为 0.25 不同样本量的四种分类方法的回判正确率盒子图。图 6 的坐标轴、盒子图与分类方法的对应关系同图 2。图 6 显示，随着样本量增大，回判正确率波动性减小，判别分析的回判正确率中位数小，波动性小。第一组合法的回判正确率中位数高于其它方法。Logistic 回归与第二组合法的回判正确率中位数接近。

(4) 解释变量个数 $p=10$ 的模拟结果

表 4 给出了解释变量个数 $p=10$ 的模拟结果，结构同表 1。表 4 显示，随着方差 σ^2 减小，四种方法的回判正确率都增大，增大得幅度小于 $p=5$ 的情况。当 $\sigma^2=1$ 和 $n=50$，第二组合法的回判正确率均值最高，为 0.8903；第一组合法的最优频数最大，为 102，第二组合法的最优频数次大，为 101。对于表 4 的其余情况，第一组合法的回判正确率均值最大，最优频数最大。当 $\sigma^2=0.25$ 且 $n=50$、100、150 和当 $\sigma^2=0.04$、0.01，第一组合法的最优频数超过了模拟次数的 90%。第二组合法的回判正确率均值和最优频数接近于 Logistic 分类，判别分析的回判正确率均值和最优频数都相对最小。

图 7 给出了样本量为 200 不同方差的 4 种方法的回判正确率盒子图。图 7 的坐标轴、盒子图与分类方法的对应关系同图 1。图 7 显示，随着 σ^2 增大，4 种分类法的回判正确率都减小。当 $\sigma^2=1$，4 种方法的回判正确率中位数的差异小，判别分析、Logistic 回归和第二组合法的回判正确率波动性差异小，第一组合法的回判正确率中位数最大，波动性稍大。当

$\sigma^2 = 0.01$、0.04、0.25，判别分析回判正确率的波动性小；第二组合法的回判正确率中位数和波动性接近于 Logistic 回归；第一组合法的回判正确率中位数较大，波动性小。

表 4 $p = 10$ 时四种方法的回判正确率和最优频数

σ^2	50	100
1	(0.8826, 0.8720, 0.8891, 0.8903)	(78 50 102 101)
	(0.9322, 0.9257, 0.9551, 0.9413)	(41 36 154 70)
	(0.9168, 0.9100, 0.9291, 0.9224)	(43 17 121 69)
	(0.9143, 0.9105, 0.9238, 0.9206)	(54 18 122 64)
0.25	(0.9745, 0.9988, 1.0000, 0.9993)	(53 195 200 196)
	(0.9531, 0.9813, 0.9916, 0.9838)	(4 133 194 134)
	(0.9502, 0.9683, 0.9811, 0.9702)	(10 77 181 78)
	(0.9475, 0.9597, 0.9721, 0.9621)	(8 46 166 56)
0.04	(0.9749, 1.0000, 1.0000, 1.0000)	(51 200 200 200)
	(0.9646, 0.9993, 1.0000, 0.9991)	(6 195 200 195)
	(0.9605, 0.9973, 0.9991, 0.9974)	(2 183 200 184)
	(0.9607, 0.9934, 0.9966, 0.9932)	(0 153 197 147)
0.01	(0.9811, 1.0000, 1.0000, 1.0000)	(71 200 200 200)
	(0.9670, 1.0000, 1.0000, 1.0000)	(6 200 200 200)
	(0.9611, 0.9996, 0.9997, 0.9995)	(1 198 200 198)
	(0.9623, 0.9996, 0.9999, 0.9996)	(0 195 200 195)

http://www.ikeen.com

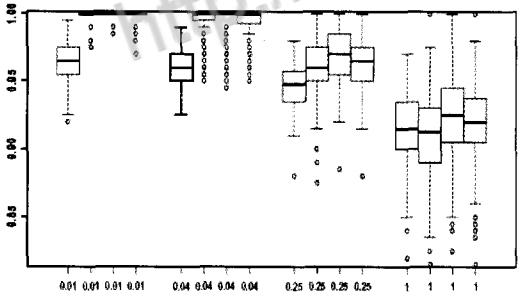

图 7 $p = 10$ 时方差不同时四种分类方法回判正确率的盒子图

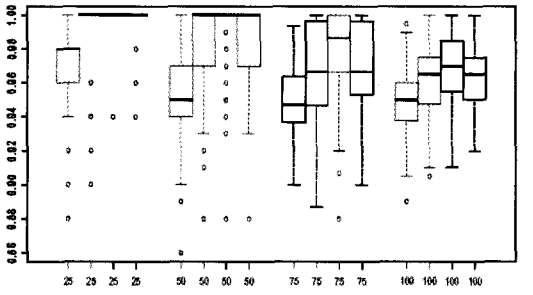

图 8 $p = 10$ 时样本量不同时四种分类方法回判正确率的盒子图

图 8 给出了方差为 0.25 不同样本量的四种分类方法的回判正确率盒子图。图 8 的坐标轴、盒子图与分类方法的对应关系同图 2。图 8 显示，随着样本量增大，回判正确率波动性变化不大，判别分析的回判正确率中位数小，波动性小；第一组合法的回判正确率中位数高于其它方法，波动性也较小；第二组合法的回判正确率的中位数和离散程度与 Logistic 回归接近。

综上所述，模拟结果显示，第一组合法和第二组合法在很多情况下都可能给出更好的分类结果，对数据进行不同解释，可能为实际研究提供有价值的参考。

3 结论

在很多情况下，新分类方法可能给出更好的分类结果，对数据进行不同解释，为实际研究提供有价值的参考。本文针对二分类问题，试图将判别分析和 Logistic 回归两种常用方法组合再进行分类，并利用随机模拟的方法研究组合分类法的回判正确率。从模拟结果显示，回判正

确率主要受到随机误差方差、变量个数、样本容量等因素的影响。在多数情况下，第一组合法的回判正确率较高。第二组合法 (基于 Logistic 回归的判别分类方法) 的回判正确率与 Logistic 回归接近。本文针对正态分布数据进行模拟，有关非正态分布情况以及函数和的最优设定等理论研究还需要进一步深入。

[参考文献]

[1] 范书山, 吕昭举, 赵守国, 陈建中, 李霞, 张瑾. 全胃肠外营养中心静脉导管感染危险因素 Logistic 回归分析 [J]. 中华医院感染学杂志, 2006, 1: 29–32.

[2] 朱燕波, 王琦, 吴承玉, 庞国明, 赵健雄, 沈世林, 夏仲元, 闫雪等. 18805 例中国成年人中医体质类型与超重和肥胖关系的 Logistic 回归分析 [J]. 中西医结合学报, 2010, 11: 1023–1028.

[3] 郭万越. 非酒精性脂肪肝相关因素 Logistic 回归分析 [J]. 实用肝脏病杂志, 2011, 5: 376–378.

[4] 田恒宇, 周汉新, 鲍世韵, 郑锦锋, 张卓, 余小舫. 胆总管结石相关因素及指标的 Logistic 回归判别分析 [J]. 中国普通外科杂志, 2007, 5: 483–485.

[5] Zavgren Christine V. Assessing the vulnerability to failure of American industrial firms [J]. Journal of Business Finance and Accounting, 1985, 12(1): 19–45.

[6] Suk Hun Lee, Jorge L Urrutia. Analysis and prediction of insolvency in the property-liability insurance industry: A comparison of logit and hazard models[J]. The Journal of Risk and Insurance, 1996, (63): 121–130.

[7] 王春峰, 万海晖, 张维. 商业银行信用风险评估及其实证研究 [J]. 管理科学学报, 1998, 1: 70–74.

[8] 梁琪. 企业经营管理预警: 主成分分析在 logistic 回归方法中的应用 [J]. 管理工程学报, 2005, 1: 100–103.

[9] 龚承刚. Logistic 回归在企业竞争力评价中的应用 [J]. 统计与决策, 2008, 19: 157–159.

[10] 缪瑾, 缪柏其, 戴小莉. 作弊因素的统计分析 [J]. 数理统计与管理, 2005, 24(3): 69–75

[11] 雷庆祝, 刘诗茂. 传统粉笔字教学与多媒体教学的影响分析 [J]. 数理统计与管理, 2011, 30(5): 942–950

[12] 陈磊, 任若恩. 时间序列判别分析技术和指数加权移动平均控制图模型在公司财务危机预警中的应用 [J]. 系统管理学报, 2009, 3: 241–248+260.

[13] 张阔, 李桂华, 李燕飞. 我国城市消费者寿险购买行为的影响因素及预测 [J]. 数理统计与管理, 2011, 30(2): 291–298.

[14] 张初兵, 高康, 杨贵军. 判别分析与 Logistic 回归的模拟比较 [J]. 统计与信息论坛, 2010, 1: 19–25.

[15] 张润楚. 多元统计分析 [M]. 北京: 科学出版社, 2006.

[16] 王学仁, 王松桂. 实用多元统计分析 [M]. 上海: 上海科学技术出版社, 1990.

[17] 张尧庭, 方开泰. 多元统计分析引论 [M]. 北京: 科学出版社, 1983.

[18] Agresti A. Categorial Data Analysis (2nd Edition) [M]. John Wiley & Sons, 2002.


知网查重限时 7折 最高可优惠 120元
本科定稿，硕博定稿，查重结果与学校一致
立即检测

免费论文查重：http://www.paperyy.com
3亿免费文献下载：http://www.ixueshu.com
超值论文自动降重：http://www.paperyy.com/reduce_repetition
PPT免费模版下载：http://ppt.ixueshu.com

阅读此文的还阅读了：

26. L_(1/2)正则化Logistic回归

27. 基于logistic回归组合预测的疾病诊断研究

28. 权核Logistic回归模型的分类和特征选择算法

29. 商业银行信用评级中逻辑回归与判别分析的对比

30. 何谓“logistic回归分析”

31. 应用SPSS软件进行多分类Logistic回归分析

32. 基于Logistic回归判别法对大学生挂科的预测

33. 逻辑回归和判别分析在财务危机预警模型中的应用

34. Logistic回归判别模型判别HIV抗体不确定者转归的可行性

35. 一种集成logistic回归与支持向量机的判别分析规则

36. Logistic回归模型和判别分析方法的比较分析

37. 基于Logistic回归和后验概率SVM的住房贷款组合评估模型

38. 如何使用SPSS对Logistic回归中分类变量进行处理

39. Logistic回归模型和判别分析方法的比较分析

40. 膨胀土的判别与分类

41. 基于Logistic回归的投资组合配置分析

42. 从线性回归到Logistic模型

43. 二分类Logistic回归插补法及其应用

44. 组织内个人知识共享的判别分析和Logistic回归分析

45. 基于logistic回归组合预测的疾病诊断研究

46. 使用Logistic回归模型进行中文文本分类

47. logistic回归诊断

48. 基于Logistic回归的数据分类问题研究

49. Logistic回归模型在判别分析中的应用

50. 基于判别分析与Logistic回归组合模型的蠓虫分类方法