基于 Gram-Schmidt 过程的判别变量筛选方法

王惠文 陈梅玲 Gilbert Saporta

(北京航空航天大学 经济管理学院, 北京 100191) (国立巴黎工艺技术学院, 巴黎 75141)

摘 要: 利用 Gram-Schmidt 过程,在自变量集合中选择对判别分类解释性最强的信息,删除对分类无显著解释作用的信息以及重复解释的信息,并把挑选出来的解释变量集合变换成若干直交变量. 一方面实现了判别分析模型中的变量筛选,同时也解决了自变量多重共线条件下的有效建模问题. 在选入变量的过程中运用 F 统计量检验变量的判别作用,更容易被统计应用人员所接受. 为了说明所提算法的合理性和有效性,以 Fisher 判别分析建模为例,通过仿真数据建模取得了合理准确的分析结论.

关 键 词: Gram-Schmidt 正交变换; 判别分析; 变量筛选; 多重相关性

中图分类号: O 212.4

文献标识码: A **文 章 编 号**: 1001-5965(2011)08-0958-04

Variable selection in discriminant analysis based on Gram-Schmidt process

Wang Huiwen Chen Meiling

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Gilbert Saporta

(Conservatoire National DesArts etMétier, Paris 75141, France)

Abstract: A new linear discriminant analysis modeling method based on Gram-Schmidt process was introduced, which firstly selected the most effective variables for classification in the independent variables set. In the meantime, the insignificant variables and the redundant information were identified and removed from the independent variables set. The selected variables were transformed into a set of orthogonal vectors by Gram-Schmidt process. Not only can the proposed method accomplish variable selection in linear discrimination, but also overcome the multi-collinearity problem effectively. Since F-statistic works as a criterion to verify the discrimination effect of each selected variable, it helps analysts to understand the analysis result. In order to test the reasonableness and effectiveness of the method, a simulation experiment was carried out. The result indicates that the proposed method can lead to a reasonable and precise conclusion.

Key words: Gram-Schmidt orthogonal transformation; discriminant analysis; variable selection; multiple correlation

判别分析是多元统计和数据挖掘中应用最为广泛的一个技术,其主要目的是通过建立模型来识别个体的所属类别. 然而在实际建模工作中,为了不遗漏重要的解释信息,人们初始经常会罗列很多的变量. 这样形成的变量集合不仅维数过高,而且蕴含的信息也非常复杂,其中包含部分没有解释意义的信息以及大量的冗余信息. 同时这样的自变量集合中经常存在比较严重的多重相关性,使得分析方法失效,因此对进入模型的自变量集合进行精心选择是十分必要的.

经典的多元线性回归或 Fisher 判别模型中可以采用前进法、后退法或逐步法进行变量筛选,然而当自变量之间高度相关时,这些方法的精度和有效性都会受到影响. 分类回归树(CART, Classi-

收稿日期: 2010-04-13

基金项目: 国家自然科学基金资助项目(70771004, 71031001, 70821061)

作者简介: 王惠文(1957 -),女,河北玉田人,教授,wanghw@ vip. sina. com.

fication and Regression Tree)在建立决策树的过程中,也会对最具有解释意义的变量自然进行筛选. 但由于该算法在每一次选择变量过程中,都必须对原始变量集合进行全部扫描,因此当原始选择的变量集合过于庞大时,建立决策树的计算工作量是比较沉重的.

近年来,随着信息技术的发展,在模式识别和机器学习等领域,人们经常要面对成千上万的变量,这使得"数据降维"成为这些领域的热点研究问题. 在众多的"数据降维"方法中,对 Gram-Schmidt 正交化过程使用以其在信息的分解、筛选和正交化等方面的优势受到人们的重视. 文献[1]首先提出采用 Gram-Schmidt 过程对线性模型中的变量进行排序,使建模过程中的变量选择的计算量从 2^p 降到 p. 该方法首先被用于机器学习和径向基函数(RBF, Radial Basis Function)网络建模[2],之后被应用于人工神经网络[3]、小波网络[4],后来又被应用于匹配追踪[5]. 文献[6]通过定义指针变量(probe feature),给出这类变量筛选的终止准则. 文献[7-8]将 Gram-Schmidt 正交过程运用于经典的 Fish 线性判别分析中提取特征过程中向量矩阵的构造,并且将其应用于人脸的识别. 文献[9]将 Gram-Schmidt 过程与经典的 t-检验相结合,给出多元回归建模过程中的变量筛选方法.

本文将讨论在判别分析的建模过程中,如何利用 Gram-Schmidt 正交化过程,在自变量集合中选择对分类判别的能力最强的信息,删除对分类无显著解释作用的信息以及重复解释的信息,并把挑选出来的解释变量集合变换成若干直交变量. 这样,一方面可以实现判别分析模型的变量筛选,同时也解决了自变量多重共线条件下的有效建模问题.

1　方法原理

Gram-Schmidt 正交化过程是线性代数中基本算法之一,对于任意一组秩为 $s \leqslant p$ 的变量集合 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p\}$,经过 Gram-Schmidt 正交变换后,得到 $Z=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p\}$. 其中存在 s 个变量是相互直交的,而另外 $(p-s)$ 个变量为零向量. 这时把 $\boldsymbol{z}_j(j=1,2,\cdots,p)$ 称为 Gram-Schmidt 变量;而把 $\boldsymbol{z}_j$ 所对应的原变量 $\boldsymbol{x}_j$ 称为与 $\boldsymbol{z}_j$ 关联的变量,可以证明经过 Gram-Schmidt 正交化过程并不会破坏原始变量集合的分布假设,而且存在 Gram-Schmidt 过程的反变换公式 $\boldsymbol{Z}=\boldsymbol{X}\ \widetilde{\boldsymbol{B}}^{-1}$, $\widetilde{\boldsymbol{B}}$ 是 Gram-Schmidt 变换过程中的系数矩阵. 相关的证明可以在文献[9]中找到,不再赘述.

以下提出基于 Gram-Schmidt 过程的判别变量筛选. 在判别分析中,设有 k 类多元正态总体,若记 y 为分类变量,y 是一个 k 类的分类变量. 特别的,当总体被分为 2 类时,y 是一个哑变量,取值为(0,1). 自变量集合为 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p\}$,若第 j 类样本数为 $n_j(j=1,2,\cdots,k)$,则总样本容量为 $n=\sum_{j=1}^{k} n_j$. 需要注意的是:为了以尽可能少的自变量实现对因变量 y 的解释与建模,在实施 Gram-Schmidt 正交化的过程中,将采用向前法的思想,根据 Gram-Schmidt 变量对分类的贡献大小进行变量筛选.

1) 设 y 是一个 k 类的分类变量. 对自变量集合 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p\}$ 做中心化处理,使每一个变量的均值为 0;为了记号方便起见,不妨记中心化后的自变量集合依然为 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p\}$.

2) 令 $\boldsymbol{z}_j^1=\boldsymbol{x}_j, j=1,2,\cdots,p$. 根据分类变量 y 分别计算 $\boldsymbol{z}_j^1$ 的组内离差平方和 E_j^1 和组间离差平方和 S_j^1. 选择 $F_j^1=\dfrac{S_j^1/(k-1)}{E_j^1/(n-k)}$ 值最大的 Gram-Schmidt 变量首先进入模型,其中 F_j^1 服从自由度为 $k-1$ 和 $n-k$ 的 F 分布. 为了记号方便起见,不妨设第 1 个被选中的变量为 $\boldsymbol{z}_1^1$,因此有 $\boldsymbol{z}_1=\boldsymbol{z}_1^1=\boldsymbol{x}_1$ 和 $F_1^1=\dfrac{S_1^1/(k-1)}{E_1^1/(n-k)}$.

在选入变量的过程中,应该考查所选入的变量是否具有判别的作用. 在经典的逐步判别中,通常考察威尔克斯统计量 Λ. 可以证明在一元的情形下,威尔克斯统计量 Λ 和 F 检验统计量是等价的[10]. 即假设 H_0: $E[\boldsymbol{x}_i \mid y=l]=E[\boldsymbol{x}_i \mid y=j], l\neq j$. 其中 $l,j=1,2,\cdots,k; i=1,2,\cdots,p$. 因此,只须考察上述的 $F^1_{1(k-1,n-1)}$ 是否能在显著性水平 α 下拒绝 H_0(H_0:无间类差异). 如果不显著,则表明一个变量也选不中;如显著,则进入下一步.

3) 令

$$\boldsymbol{z}_j^2=\boldsymbol{x}_j-\frac{\boldsymbol{x}_j'\boldsymbol{z}_1}{\boldsymbol{z}_1'\boldsymbol{z}_1}\boldsymbol{z}_1 \qquad j=2,3,\cdots,p \qquad (1)$$

分别计算 $\boldsymbol{z}_j^2(j=2,3,\cdots,p)$ 的组内离差平方和 E_j^2 和组间离差平方和 S_j^2. 选择 $F_j^2=\dfrac{S_j^2/(k-1)}{E_j^2/(n-k)}$ 值最大的 Gram-Schmidt 变量进入模型. 不妨设为 $\boldsymbol{z}_2=\boldsymbol{x}_2-\boldsymbol{b}_{21}\boldsymbol{z}_1$,其中 $\boldsymbol{b}_{21}=\dfrac{\boldsymbol{x}_2'\boldsymbol{z}_1}{\boldsymbol{z}_1'\boldsymbol{z}_1}$. 第 2 个被选中的关联变量为 $\boldsymbol{x}_2$.

4) 令

$$z_j^3 = x_j - \frac{x_j' z_1}{z_1' z_1} z_1 - \frac{x_j' z_2}{z_2' z_2} z_2 \qquad j = 3,4,\cdots,p \tag{2}$$

分别计算 $z_j^3 (j=3,4,\cdots,p)$ 的组内离差平方和 E_j^3 和组间离差平方和 S_j^3. 选择 $F_j^3 = \frac{S_j^3/(k-1)}{E_j^3/(n-k)}$ 值最大的 Gram-Schmidt 变量进入模型. 不妨设为 $z_3 = x_3 - b_{31} z_1 - b_{32} z_2$, 其中 $b_{31} = \frac{x_3' z_1}{z_1' z_1}, b_{32} = \frac{x_3' z_2}{z_2' z_2}$. 则第 3 个关联变量为 x_3.

重复上面的步骤, 直到判别模型外边所有的变量经过 Gram-Schmidt 处理后都不能通过 F 检验, 此时模型外的所有 Gram-Schmidt 变量均不能提供附加信息. 根据最终选择出来的 Gram-Schmidt 变量 $z_1, z_2, \cdots, z_m (m \leqslant p)$, 建立判别分析模型. 由于 Gram-Schmidt 变量是相互正交的, 所以可以避免多重共线对建模过程的影响. 例如, 采用 $z_1, z_2, \cdots, z_m$ 建立了 Fisher 判别模型如下:

$$F = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_m z_m \tag{3}$$

根据 Gram-Schmidt 过程的反变换公式, 还可以将 y 关于 $z_1, z_2, \cdots, z_m$ 的 Fisher 判别模型变换成关于原始自变量 $x_1, x_2, \cdots, x_m$ 的模型:

$$F = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_m x_m \tag{4}$$

从本节所给出的变量筛选与建模方法可以看出, 对判别分类更有解释意义的信息被挑选出来, 重复信息在 Gram-Schmidt 过程中被减掉, 而对分类无解释意义的信息也被排除. 此外, 由于所使用的 Schmidt 变量不存在相关性, 因此可以有效地克服自变量多重相关所造成的算法失效问题.

2 仿真案例研究

本节将采用仿真数据进行研究, 以验证本文中提出的基于 Gram-Schmidt 过程的判别变量筛选方法. 首先, 生成自变量集合为 x_1, x_2 的 2 类总体二维正态分布随机数, 各类的样本容量分别为 100, 因此总样本容量为 $n = 100 + 100 = 200$. 此时 y 的取值为 (0,1). 2 类随机数的数字特征如表 1 所示.

表 1 2 类总体的二维正态分布随机数的数字特征

类别	$E(x_1)$	$E(x_2)$	$D(x_1)$	$D(x_2)$	r
0	1	3	1	1	−0.14
1	3	1	1	1	−0.19

进一步, 引进变量 $x_j, j=3,4,\cdots,8$, 其中 $x_j = x_2 + \varepsilon_j, j=3,4,5,6; x_7 = 0.9x_1 + 0.1x_2 + \varepsilon_7; x_8 = 0.2x_1 + 0.8x_2 + \varepsilon_8; \varepsilon_j (j=3,4,\cdots,8)$ 服从 (−0.1, 0.1) 的均匀分布, 这 6 个变量均与 x_1, x_2 高度相关.

下面给出基于 Gram-Schmidt 过程的判别变量筛选方法和建模过程, 首先对自变量集合做中心化处理, 将处理后的自变量记为 $z_j^1, j=1,2,\cdots,8$. 然后根据分类变量 y 分别计算 $z_j^1 (j=1,2,\cdots,8)$ 的 $F_j^1 = \frac{S_j^1/(k-1)}{E_j^1/(n-k)}, j=1,2,\cdots,8$, 结果见表 2.

表 2 判别变量 z_j^1 的 F 检验统计量

变量	z_1^1	z_2^1	z_3^1	z_4^1
F 检验统计量	164.5	236.9	232.0	236.2
变量	z_5^1	z_6^1	z_7^1	z_8^1
F 检验统计量	234.2	232.9	205.2	232.7

自由度为 1 和 198 的 F 检验统计量的在置信水平为 5% 的阈值查表为 3.91. 选择 F 检验统计量绝对值最大的自变量为 z_1, 即 $z_1 = z_2^1$. 分别将 $z_1^1, z_3^1, z_4^1, z_5^1, z_6^1, z_7^1, z_8^1$ 与 z_1 做 Gram-Schmidt 正交变换 $z_j^2, j=1,3,4,5,6,7,8$. 然后根据分类变量 y 分别计算 $z_j^2 (j=1,3,4,5,6,7,8)$ 的 $F_j^2 = \frac{S_j^2/(k-1)}{E_j^2/(n-k)} (j=1,3,4,5,6,7,8)$, 结果如表 3 所示.

表 3 判别变量 z_j^2 的 F 检验统计量

变量	z_1^2	z_3^2	z_4^2	z_5^2	z_6^2	z_7^2	z_8^2
F 检验统计量	15.2	0.049	0.387	0.253	0.008	14.8	0.002

可以看到经过 Gram-Schmidt 正交变换后, 选择 F 检验统计量最大的 z_1^2 的进入模型, 即 $z_2 = z_1^2$. 同时变换后的 5 个变量 $z_3^1, z_4^1, z_5^1, z_6^1, z_8^1$ 因为其 F 检验统计量已经远远小于阈值, 说明经过变换扣除分类的重复解释信息, 已经不再具有判别效力了, 可以直接删除. 下一步让 z_7^2 与 z_2 做 Gram-Schmidt 正交变换得到 z_7^3, 计算 z_7^3 对应的 $F_7^3 = 1.57$, 此时 z_7^3 的 F 检验统计量已经小于阈值, 不能进入模型.

因此, 根据最终选择出来的 Gram-Schmidt 变量 z_1, z_2, 本例中采用 z_1, z_2 建立了 Fisher 判别模型, 得到如下建模结果:

$$F = 1.525 z_1 - 0.771 z_2$$

判别分析的结果如表 4.

根据 Gram-Schmidt 过程的反变换公式 (4), 还可以将 y 关于自变量 x_1, x_2 的模型:

$$F = -0.353 - 0.578 x_1 + 0.764 x_2$$

图 1 和图 2 分别为仿真样本的原始分类图和

Fisher 判别分析后的结果图,可以看到,采用 2 个变量已经可以很好地对样本进行分类. 同时在建模的过程中,可以很好地挑选出对判别分类更有解释意义的信息,重复信息则通过 Gram-Schmidt 变换被排除.

表 4　Fisher 判别结果

原始分类	建模分类结果		总计(样本个数)
	0	1	
0	90	10	100
1	9	91	100

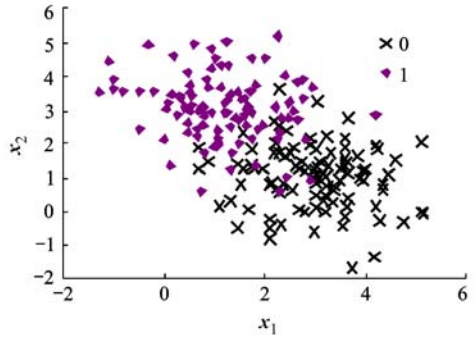

图 1　仿真数据的实际分类图

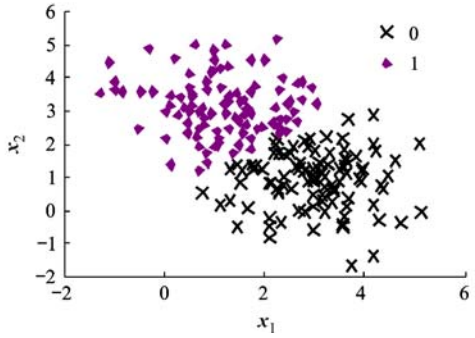

图 2　判别分析的结果图

3　结　论

在实际工作中,根据问题本身的专业理论和有关经验,研究人员罗列出可能与判别相关的变量往往很多,在这样形成的自变量集合中,虽然不遗漏重要的解释信息,但是包含一部分没有解释作用的信息和大量的冗余信息(即重复解释的变量). 如果直接采用所有这些自变量做判别分析,一方面不符合统计模型的参数节省原则,而且由于自变量集合中包含多重相关性,还将对模型结论造成不良影响. 所以长期以来,如何更加合理有效地筛选解释变量一直是许多应用人员非常关心的问题.

本文提出利用 Gram-Schmidt 过程,在自变量集合中选择分类判别能力最强的变量,并且在变量筛选过程中,把对因变量没有显著解释作用的信息以及冗余的解释信息有效地分解出来并且排除掉. 这样一来,使用 Gram-Schmidt 变量进行建立判别模型,不仅可以有效克服自变量集合多重共线对回归建模的不良影响,并且由于其信息分解结构清晰,使对模型的解释也更加容易,有效地解决了大规模高维数据的判别问题. 此外,由于该模型除了利用 Gram-Schmidt 过程进行信息提取以外,在建模的过程中运用的 F 检验统计量简单直接,更容易被统计应用人员所接受.

参考文献(References)

[1] Chen S, Billings S A, Luo W. Orthogonal least squares methods and their application to non-linear system identification[J]. International Journal of Control, 1989, 50(5): 1873 – 1896

[2] Chen S, Cowan C F N, Grant P M. Orthogonal least squares learning algorithm for radial basis function networks[J]. IEEE Transaction on Neural Networks, 1991, 2(2): 302 – 309

[3] Urbani D, Roussel-Ragot P, Personnaz L, et al. The selection of neural models of nonlinear dynamical systems by statistical tests [C]//Vlontzos J, Hwang J, Wilson E. Neural Networks for Signal Processing IV. Piscataway, NJ: IEEE, 1994: 229 – 237

[4] Oussar Y, Dreyfus G. Initialization by selection for wavelet network training[J]. Neurocomputing, 2000(34): 131 – 143

[5] Vincent P, Bengio Y. Kernel matching pursuit [J]. Machine Learning, 2001(48): 165 – 187

[6] Stoppiglia H, Dreyfus G, Dubois R, et al. Ranking a random feature for variable and feature selection[J]. The Journal of Machine Learning Research, 2003(3): 1399 – 1414

[7] Zheng Wenming, Zou Cairong, Zhao Li. Real-time face recognition using Gram-Schmidt orthogonalization for LDA[C]//Proceedings-International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2004: 403 – 406

[8] He Yunhui. Modified generalized discriminant analysis using kernel Gram-Schmidt orthogonalization in difference space for face recognition[C]//Proceedings 2009 2nd International Workshop on Knowledge Discovery and Data Mining. Piscataway, NJ: IEEE, 2009: 36 – 39

[9] 王惠文,陈梅玲,Gilbert Saporta. Gram-Schmidt 回归及在刀具磨损预报中的应用[J]. 北京航空航天大学学报,2008,34(6):729 – 733
Wang Huiwen, Chen Meiling, Gilbert Saporta. Gram-Schmidt regression and application in cutting tool abrasion prediction[J]. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(6): 729 – 733(in Chinese)

[10] Johnson R A, Wichern D W. Applied multivariate statistical analysis[M]. 6th ed. Beijing: Tsinghua University Press, 2008

(编　辑:文丽芳)