文章编号:1000-6788(2005)07-0131-05

Bayes-可拓判别

李日华,毛　凯,周　刚
(海军航空工程学院基础部,山东 烟台 264001)

摘要:　研究了 Bayes 判别与可拓判别两种判别法的优缺点,在 Bayes 判别规则中引入了关联度,建立了集 Bayes 判别与可拓判别两种判别方法优点的 Bayes-可拓判别法,并证明了 Bayes 判别法与可拓判别法都是 Bayes-可拓判别法在一定条件下的特例. 通过对举例计算结果的讨论,显示了 Bayes-可拓判别法的优越性.

关键词:　Bayes 判别;可拓学;判别法;关联度
中图分类号:　N94;O21　　**文献标识码**:　A

Bayes-Extension Discrimination

LI Ri-hua, MAO Kai, ZHOU Gang
(Department of Basic Science, NAEI, Yantai 264001, China)

Abstract:　After discussed the advantages and disadvantages of the Bayes discrimination method and the extension discrimination method, and then introduced the interrelating degree into Bayes discrimination rule, the authors of this paper have obtained a new Bayes-Extension discrimination method, which combines the advantages of the former two methods, and also proved that the former two methods are just the special cases of the new method under a certain condition. At the end of this paper, the discussion of the computing results have disclosed clearly the advantages of the Bayes-Extension method.

Key words:　Bayas discrimination; extension; discrimination method; interrelating degree

1　引言

Bayes 判别法[1]是判别分析中一种常用的重要判别方法,它是在考虑了各总体出现的先验概率及误判损失的前提下,用取得的样本修正先验概率分布,进而判别样本(样品)的归属. 而可拓学中的可拓判别法[2]及其应用[3]是从实际问题出发,考虑到决定一个样品特性的各个指标往往不是相互独立的,而是具有某种内在的联系,并且它们对总体特性影响的大小各不相同,通过建立各总体和待判样品的物元模型,利用经典域、节域计算出待判样品(物元)与各总体(物元)的关联度,样品物元与某总体物元的关联度越大,则它们的符合程度就愈佳,故应将样品判属与其关联度最大的总体. 本文将 Bayes 判别法与可拓判别法相结合得到一种 Bayes-可拓判别法,它集两种判别方法的优点,考虑问题更全面,判别将更准确. Bayes-可拓判别法是多指标参数的可拓概率[4]统计计算判别模型,其判别计算结果完全是一种数量结果,具有很强的可操作性.

2　Bayes 判别

设有 m 个总体 $G_i \sim f_i(X)$,$(i = 1, 2, \cdots, m)$,它们的先验概率分别为 $q_1, q_2, \cdots, q_m$,显然 $q_i \geq 0$,$\sum_{i=1}^{m} q_i = 1$. 若将属于总体 G_i 的样品错判给总体 G_j 时,造成的误判损失记为 $c(j|i)$,显然 $c(j|i) \geq 0$,且 $c(i|i) = 0$,对任意的 $i, j = 1, 2, \cdots, m$ 成立.

在 Bayes 判别规则 $R = \{R_1, R_2, \cdots, R_m\}$ 下,将属于 G_i 的样品错判给 G_j 的误判概率记为 $P(j|i, R)$

收稿日期:2004-06-25
作者简介:李日华(1955-),男,现任海军航空工程学院基础部数学教研室教授.

$= \int_{\Omega_j} f_i(X)\,\mathrm{d}X$, $(i, j = 1, 2, \cdots, m, i \neq j)$. 因此, 规则 R 把来自总体 G_i 的样品 X 错判至其它总体的平均损失按误判概率加权平均为

$$r(i, R) = \sum_{j=1}^{m} [c(j \mid i) \cdot P(j \mid i, R)].$$

因而 Bayes 判别规则 R 就是要选择 $R_1, R_2, \cdots, R_m$, 使总的平均损失

$$g(R) = \sum_{i=1}^{m} q_i \sum_{j=1}^{m} c(j \mid i)\, P(j \mid i, R) \tag{1}$$

达到最小. 一般有

定理 2.1 设总体 $G_i \sim f_i(X)$, 先验概率为 q_i, $(i = 1, 2, \cdots, m)$, 当误判损失为 $c(j \mid i)$ 时, 则划分 $R = \{R_1, R_2, \cdots, R_m\}$ 的 Bayes 解为

$$R_l = \{X : h_l(X) = \min_j h_j(X)\}, l = 1, 2, \cdots, m,$$

其中

$$h_j(X) = \sum_{i=1}^{m} q_i c(j \mid i) f_i(X), (j = 1, 2, \cdots, m). \tag{2}$$

证明: 参见文献[1].

虽然 Bayes 判别法考虑了误判损失与各总体的先验概率及其分布, 但它既没有考虑决定样品 X 特性的各个指标(即 X 各分量)与总体 G_i 相应指标的"贴近"程度, 也没有考虑样品的各个具体指标对样品总体特性影响的大小, 而实际问题中各指标对样品 X 的总体特性影响的大小往往是不同的.

3 可拓判别法

考虑到决定一个样品的各个指标对总体 G_i 相应指标的"贴近"程度, 以及各个指标对总体特性的影响的大小. 通过建立各总体和待判样品的物元模型, 利用经典域、节域计算出待判样品(物元)与各总体(物元)的关联度, 如果样品物元与某总体物元的关联度越大, 则它们的符合程度就愈佳, 故应将样品判属与其关联度最大的总体.

3.1 建立各总体和待判样品的物元模型

m 个总体的物元模型为

$$R_{0i} = (N_{0i}, C, V_{0i}) = \begin{pmatrix} N_{0i}, & c_1, & V_{0i1} \\ & c_2, & V_{0i2} \\ & \cdots & \cdots \\ & c_p, & V_{0ip} \end{pmatrix} = \begin{pmatrix} N_{0i}, & c_1, & \langle a_{0i1}, b_{0i1} \rangle \\ & c_2, & \langle a_{0i2}, b_{0i2} \rangle \\ & \cdots & \cdots \\ & c_p, & \langle a_{0ip}, b_{0ip} \rangle \end{pmatrix}$$

其中 N_{0i} 表示第 i $(i = 1, 2, \cdots, m)$ 个总体的名称, c_j $(j = 1, 2, \cdots, p)$ 表示 N_{0i} 的特征, V_{0ij} 为 N_{0i} 关于特征 c_j 所规定的量值范围 ——经典域.

样品 X 的节域的物元模型为

$$R_X = (X, C, V_X) = \begin{pmatrix} X, & c_1, & V_{X1} \\ & c_2, & V_{X2} \\ & \cdots & \cdots \\ & c_p, & V_{Xp} \end{pmatrix} = \begin{pmatrix} X, & c_1, & \langle a_{0X1}, b_{0X1} \rangle \\ & c_2, & \langle a_{0X2}, b_{0X2} \rangle \\ & \cdots & \cdots \\ & c_p, & \langle a_{0Xp}, b_{0Xp} \rangle \end{pmatrix}$$

于是待判样品 X 的物元模型为

$$R = \begin{pmatrix} X, & c_1, & v_1 \\ & c_2, & v_2 \\ & \cdots & \cdots \\ & c_p, & v_p \end{pmatrix}$$

由于 X 是由特征和量值所确定, 下面也记 $X = (v_1, v_2, \cdots, v_p)^{\mathrm{T}}$, 其中 v_j 为 X 关于 c_j 的量值, 即对待判

样品检测所得的具体数据.

3.2　确定各特征指标的权系数

对非满足不可的特征 c_k,记其权系数为 Λ,若待判样品 X 的相应量值 $v_k \notin V_{0lk}$,则样品 X 不属于 R_{0l} 或 G_l,$0 \leq l \leq m$,此时将总体 G_l 排除. 其它特征 c_j $(j \neq k)$ 的权系数记为 α_j,α_j $(j = 1, 2, \cdots, p, j \neq k)$,显然 $\sum_{\substack{j=1 \\ j \neq k}}^{p} \alpha_j = 1$.

3.3　确定关联度进行判别

对 $v_k \in V_{0ik}$,$i = 1, 2, \cdots, m, i \neq l$,待判样品 X 的各特征指标关于各总体的相应指标的关联度为

$$K_i(v_j) = \frac{\rho(v_j, V_{0ij})}{\rho(v_j, V_{Xj}) - \rho(v_j, V_{0ij})}. \tag{3}$$

根据距的定义(文献[2]),其中

$$\rho(v_j, V_{0ij}) = \left| v_j - \frac{a_{0ij} + b_{0ij}}{2} \right| - \frac{b_{0ij} - a_{0ij}}{2},$$

$$\rho(v_j, V_{Xj}) = \left| v_j - \frac{a_{0Xj} + b_{0Xj}}{2} \right| - \frac{b_{0Xj} - a_{0Xj}}{2},$$

$i = 1, 2, \cdots, m; j = 1, 2, \cdots, p.$

于是待判样品 X 关于总体 R_{0i} 或 G_i 的关联度

$$K_i(X) = \sum_{j=1}^{p} \alpha_j K_i(v_j), (i = 1, 2, \cdots, m, i \neq l). \tag{4}$$

如果

$$K_{i_0}(X) = \max_{i_0 \in \{1, 2, \cdots, m\}} K_i(X), \tag{5}$$

则将样品判属 G_{i_0}.

虽然可拓判别法考虑了 Bayes 判别法所没有考虑的两点,但这里没有考虑可能产生的误判损失以及总体 G_i 出现的先验概率和分布.

4　Bayes-可拓判别法

考虑到在 Bayes 判别规则 R 中,选择 $R_1, R_2, \cdots, R_m$,使总的平均损失

$$g(R) = \sum_{i=1}^{m} q_i \sum_{j=1}^{m} c(j \mid i)\, P(j \mid i, R)$$

达到最小. 结合由(4)式所确定的关联度 $K_i = K_i(X)$ $(i = 1, 2, \cdots, m)$,则所寻找的判别规则 R 应使误判平均损失

$$S(R) = \sum_{i=1}^{m} q_i K_i \sum_{j=1}^{m} c(j \mid i)\, P(j \mid i, R) \tag{6}$$

达到最小. 这时有

定理 4.1　设总体 $G_i \sim f_i(X)$,先验概率为 q_i,$(i = 1, 2, \cdots, m)$,误判损失为 $c(j \mid i)$,样品 X 关于 G_i 的关联度由(4)式确定时,则划分 $R = \{R_1, R_2, \cdots, R_m\}$ 的 Bayes-可拓解为

$$R_l = \{X : h_l(X) = \min_j h_j(X)\}, l = 1, 2, \cdots, m,$$

其中

$$h_j(X) = \sum_{i=1}^{m} q_i K_i c(j \mid i)\, f_i(X). \tag{7}$$

证　由(6)式

$$S(R) = \sum_{i=1}^{m} q_i K_i \sum_{l=1}^{m} c(l \mid i)\, P(l \mid i, R) = \sum_{i=1}^{m} q_i K_i \sum_{l=1}^{m} c(l \mid i) \int_{l} f_i(X)\, \mathrm{d}X$$

$$= \sum_{l=1}^{m} \int_{R_l} \sum_{i=1}^{m} q_i K_i c(l \mid i) f_i(X) \mathrm{d}X = \sum_{l=1}^{m} \int_{R_l} h_l(X) \mathrm{d}X.$$

要使 $S(R)$ 达到最小,等价于在 R_l 上 $h_l(X)$ 为所有 $h_j(X)$ 中最小者. 证毕

推论 4.1 若误判损失是 $c(j \mid i) = 1, (i \neq j), c(i \mid i) = 0$ 时,则划分 $R = \{R_1, R_2, \cdots, R_m\}$ 的 Bayes-可拓解为

$$R_l = \{X : q_l K_l f_l(X) = \max_j q_j K_j f_j(X)\}. \tag{8}$$

证 由(7)式

$$h_j(X) = \sum_{i=1}^{m} q_i K_i c(j \mid i) f_i(X) = \sum_{\substack{i=1 \\ i \neq j}}^{m} q_i K_i f_i(X)$$

$$= \sum_{i=1}^{m} q_i K_i f_i(X) - q_j K_j f_j(X) = P(X) - q_j K_j f_j(X),$$

其中 $P(X)$ 是与 j 无关的常量,因而(7)式中要使 $h_l(X)$ 取到 $h_j(X)$ $(j = 1, 2, \cdots, m)$ 中的最小值,等价于 $q_l K_l f_l(X)$ 在所有 $q_j K_j f_j(X)$ $(j = 1, 2, \cdots, m)$ 中取到最大值. 证毕

(8)式的意义是非常明显的,即当判对了无损失,判错了误判损失都相同,且样品 X 关于总体 G_l 的关联度与 $q_l f_l(X)$ 的乘积最大时,当然应把样品 X 判属总体 G_l.

由定理 4.1 及其推论 4.1 可知,当 $K_i = 1$ $(i = 1, 2, \cdots, m)$,即样品 X 与各总体的关联度均相同时,Bayes-可拓判别法就是 Bayes 判别法;而当 $q_j f_j(X)$ $(j = 1, 2, \cdots m)$ 均相同,且 $c(j \mid i) = 1, (i \neq j), c(i \mid i) = 0$ 时,Bayes-可拓判别法也就退化为可拓判别法.

Bayes-可拓判别法就是在可拓判别法的基础上考虑了误判损失以及各总体出现的先验概率及其分布,考虑的因素更全面了,判别结论将更准确. 在具体计算时,首先使用可拓判别法计算出样品 X 与各总体的关联度 K_i $(i = 1, 2, \cdots, m)$,如果需要进一步考虑误判损失及各总体出现的先验概率和分布时,只需将 K_i 按 $q_i c(j \mid i) f_i(X)$ 加权平均,计算出各

$$h_j(X) = \sum_{i=1}^{m} q_i K_i c(j \mid i) f_i(X) \quad (j = 1, 2, \cdots, m),$$

再比较 $h_1(X), h_2(X), \cdots, h_m(X)$,选取其中最小者判定样品 X 来自该总体.

5 举例

考虑由三个特征 $(p = 3)$ 确定的三个总体 $(m = 3)$,它们的经典域由物元表示为

$$R_{01} = \begin{pmatrix} N_{01}, & c_1, & \langle 0.7, 0.85 \rangle \\ & c_2, & \langle 0.5, 0.6 \rangle \\ & c_3, & \langle 6.0, 7.8 \rangle \end{pmatrix}, \quad R_{02} = \begin{pmatrix} N_{02}, & c_1, & \langle 0.6, 0.8 \rangle \\ & c_2, & \langle 0.65, 0.8 \rangle \\ & c_3, & \langle 9.0, 11.0 \rangle \end{pmatrix},$$

$$R_{03} = \begin{pmatrix} N_{03}, & c_1, & \langle 0.85, 0.95 \rangle \\ & c_2, & \langle 0.8, 0.9 \rangle \\ & c_3, & \langle 7.8, 9.0 \rangle \end{pmatrix},$$

节域为

$$R_X = \begin{pmatrix} X, & c_1, & \langle 0.6, 0.95 \rangle \\ & c_2, & \langle 0.5, 0.9 \rangle \\ & c_3, & \langle 6.0, 11.0 \rangle \end{pmatrix},$$

样品 X 为

$$R = \begin{pmatrix} X, & c_1, & 0.8 \\ & c_2, & 0.7 \\ & c_3, & 8.8 \end{pmatrix}.$$

下面来判别它的归属.

利用关联度公式计算样品 X 关于 G_i,即 R_{0i} $(i = 1, 2, 3)$ 的关联度得:

$$K_1(v_1) = 0.5000,\quad K_1(v_2) = -0.3333,\quad K_1(v_3) = -0.3125;$$
$$K_2(v_1) = 0.0000,\quad K_2(v_2) = 0.3333,\quad K_2(v_3) = -0.0833;$$
$$K_3(v_1) = -0.2500,\quad K_3(v_2) = -0.3333,\quad K_3(v_3) = 0.1000.$$

如果取权系数 $\alpha_1 = 0.5$, $\alpha_2 = 0.3$, $\alpha_3 = 0.2$, 则样品 X 为关于 G_i, 即 R_{0i} $(i = 1,2,3)$ 的关联度分别:

$$K_1(X) = 0.08751,\ K_2(X) = 0.08333,\ K_3(X) = -0.20499.$$

由于 $K_1(X) > K_2(X) > K_3(X)$, 按可拓判别法应将样品 X 判属总体 G_1.

如果设总体 G_i $(i = 1,2,3)$ 均是协方差阵为 $V = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$ 的三元正态总体, 误判损失分别为 $C(1|2) = 10$, $C(1|3) = 1$, $C(2|1) = 12$, $C(2|3) = 2$, $C(3|1) = 12$, $C(3|2) = 2$, $c(i|i) = 0$, $(i = 1,2,3)$, 且 $q_1 = 0.3$, $q_2 = 0.4$, $q_3 = 0.3$, 下面用 Bayes-可拓判别法进行判别.

此时均值向量分别为 $\mu_1 = \begin{pmatrix} 0.775 \\ 0.55 \\ 6.9 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 0.7 \\ 0.725 \\ 10 \end{pmatrix}$, $\mu_3 = \begin{pmatrix} 0.9 \\ 0.85 \\ 8.4 \end{pmatrix}$.

经计算得样品 $X = (0.8, 0.7, 8.8)^{\mathrm{T}}$ 关于 G_i $(i = 1,2,3)$ 的密度函数分别为

$$f_1(X) = \frac{1}{(2\pi)^{3/2}\sqrt{2}}\exp(-0.8870),$$
$$f_2(X) = \frac{1}{(2\pi)^{3/2}\sqrt{2}}\exp(-0.4353),$$
$$f_3(X) = \frac{1}{(2\pi)^{3/2}\sqrt{2}}\exp(-0.06875),$$

于是有 $h_1(X) = \frac{1}{(2\pi)^{3/2}\sqrt{2}}0.1583$, $h_2(X) = \frac{1}{(2\pi)^{3/2}\sqrt{2}}0.01494$, $h_3(X) = \frac{1}{(2\pi)^{3/2}\sqrt{2}}0.1729$, 显然 $h_2(X) < h_1(X) < h_3(X)$, 故应将样品 X 判属总体 G_2.

6　结束语

将 Bayes 判别法与可拓判别法相结合得到的 Bayes-可拓判别法, 集中了两种判别方法的优点, 考虑问题更全面, 使判别结论更准确了. 特别是当各关联度都相等时, Bayes-可拓判别法就退化为 Bayes 判别法; 而由推论 4.1 知, 当不考虑误判损失, 且各总体出现的先验概率均相同又同分布时, 判别法又退化为可拓判别法. 在此意义上, Bayes 判别法与可拓判别法都是 Bayes-可拓判别法的特例.

参考文献:

[1] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1982, 194 - 215.
Zhang Yaoting, Fang Kaitai. The Introductory Theory of Multivariate Statistics[M]. Beijing: The Science Press, 1982, 194 - 215.

[2] 蔡文, 杨春燕, 林伟初. 可拓工程方法[M]. 北京: 科学出版社, 1997, 202 - 209.
Cai Wen, Yang Chunyan, Lin Weichu. Extension Engineering Method[M]. Beijing: The Science Press, 1997, 202 - 209.

[3] 蒋淳, 等. 中强地震预报综合评判物元模型及其应用[J]. 系统工程理论与实践, 1998, 18(1): 135 - 138.
Jiang Chun, et al. The matter-element model of multi-factorial evaluation and its application to earthquake prediction[J]. System Engineering - Theory & Practice, 1998, 18(1): 135 - 138.

[4] 李日华, 姜殿玉. 可拓事件与可拓概率[J]. 广东工业大学学报, 1999, 16(4): 96 - 101.
Li Rihua, Jiang Dianyu. Extension event and extension probability[J]. Journal of Guangdong University of Technology, 1999, 16(4): 96 - 101.