

Bayes 判别在巨猿分类中的应用

王 令 红 方开泰 郑玉颖

(中国科学院古脊椎动物与古人类研究所) (中国科学院应用数学研究所)

Bayes 判别分析是多元分析中的方法之一。多元分析是用数理统计的理论研究多项指标问题的一种方法,它的理论在本世纪 30 年代就得到相当完整的发展。最早用于考古学和古生物学的研究之中,后来应用面越来越广,目前已广泛地应用于地质、医学、心理学、行为科学、标准化、经济、海洋、农业等方面,并取得了很好的效果。

通过电子计算机的运算,采用数学方法进行分类,已成为当代生物分类的三个主要流派(综合学派、分支学派和表现型学派)之一,数量分类即是表现型学派据以分类的实质内容。就世界范围而言,数量分类方法已经进入了生物各大门类的分类之中,甚至于渗透到生物学的其他领域。多元分析是数量分类中目前比较理想的方法。

在人类学中,多元分析已被应用在活体测量和骨骼测量结果的统计分析、各部化石的研究和对比、人群类型的分组和人类的起源等各个方面。在我国丰富的化石材料中,巨猿和不少其他材料也早已被国外学者使用多元分析的方法研究过,并得到不同的结论。

一、问题的提出

从 1935 年以来,陆续发现了大量的巨猿化石,包括四个下颌骨和一千余枚牙齿。它们表现了既似人又似猿的特征,以及在硕大的方向上逐渐特化的性质,这引起了中外学者广泛的注意。其中,有一些外国学者从巨猿的测量特征的数量统计分析和对比的角度出发,探讨它的分类地位。

Fraye (1973) 逐项地分别对比了巨猿和粗壮南方古猿的下颌骨和牙齿的测量特征构成的几项指数。研究结果是,除个别项目以外,巨猿的其他指数都落入南方古猿的变异范围之中。因此,巨猿与南方古猿一样,是人科成员。在中国发现的步氏巨猿仅因为生存时代与南方古猿和直立人重叠才成为人科的绝灭旁支;而印度的比拉斯普巨猿则可以作为南方古猿的祖先的“最合适的人选”。

与此同时,(Robinson and Steudel, 1973 年)使用了多元判别分析的方法研究了巨猿的牙齿,并与现代人、现生三种大型类人猿、直立人和南方古猿(文中称之为“非洲人”和“傍人”)的有关材料进行对比。结论是巨猿居于人科和猿科之间的位置,总的来说,更接近于粗壮南方古猿,并可能是后者的祖先。

Corruccini (1975 年)认为单凭逐项分析经主观选择构成的指数可以使任何设想都得到证明,“只需想办法搞出几个指数来就行了。”他认为 Frayer 的单变量测验方法没有把所有特征看作为一个“统一的形态类型”,也避免不了这类错误。他仍使用 Frayer 使用的

同样材料，但以几种多元分析的不同方法对这些数据进行重新处理。其结果也就完全不同：虽然巨猿在整体型态上距离南方古猿比距离大猿更近些，但更重要的是南方古猿与大猿的接近程度远远超过了它与巨猿的关系。既然南方古猿与大猿分别属于不同的科，巨猿与南方古猿的关系应该更远，巨猿是早期就特化了一支。

多元分析对于单变量测验的优越性是显而易见的。Robinson 等对巨猿牙齿的分析，Corruccini 对巨猿下颌和牙齿的整体分析是否都概括了各自部分的形态特点？是否对于巨猿分类地位的确定都作出了贡献？

从树居的古猿下到地面，直到人类的形成，这中间经过了漫长的一个过渡时期。这个过渡期间的生物经历了生态、食性、行动和行为习惯、身体各部功能等等各方面的变化。其中必定分为若干步骤。因此，人类的特征和猿类特征不同程度上镶嵌式共存于整个过渡期间的生物之中。这种镶嵌现象在巨猿的牙齿和下颌骨上到底在多大程度上得到反映？

这些都促使我们对巨猿的牙齿和下颌骨的测量特征进行多元分析。

二、方 法

Bayes 判别分析的原理是这样的：

设已知有 k 类物体 g_1, g_2, \dots, g_k ，每一个物体均测了 m 个指标，记作 x_1, x_2, \dots, x_m 每一类的分布密度为： $P_1(x), P_2(x), \dots, P_k(x)$ 这 k 类出现的机会不一定是相等的，用 $q_1, q_2, \dots, q(k)$ 表示它们出现的概率(常称作先验概率)，显然 $q_1 > 0, q_2 > 0, \dots, q_k > 0, q_1 + q_2 + \dots + q_k = 1$ 。并根据实际问题，选择一个损失函数，用 $L(i/j)$ 表示物体本来属于 g_i ，而误判为 g_j 的损失，在实际中常采用相等损失的函数，即：

$$L(i/j) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

即把 g_i 的物体误判为 g_j 的损失，与把 g_j 的物体误判为 g_i 的损失是一样的。在我们的计算中就是采取了这个函数。

所谓 Bayes 判别分析就是设法算出判别函数与判别规则，使误判的平均损失达到极小。在相当多的实际问题中，常可假定各类的分布密度是协差阵的多维(m 维)正态分布，这时，Bayes 判别可用如下步骤来计算：

用 $x_{ij}^{(l)}$ 表示第 l 类物体中第 j 个指标的第 i 个样本的值，显然 $l = 1, 2, \dots, k; j = 1, 2, \dots, m; i = 1, 2, \dots, n_l; n_l$ 为第 l 类中的样本数。通常将数据形成一个表，表的形式为：

$$\left. \begin{matrix} x_{11}^{(1)} & x_{12}^{(1)} & \dots & x_{1m}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \dots & x_{2m}^{(1)} \\ \dots & \dots & \dots & \dots \\ x_{n_1 1}^{(1)} & x_{n_1 2}^{(1)} & \dots & x_{n_1 m}^{(1)} \\ \dots & \dots & \dots & \dots \end{matrix} \right\} g_1$$

$$\left. \begin{matrix} x_{11}^{(k)} & x_{12}^{(k)} & \dots & x_{1m}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \dots & x_{2m}^{(k)} \\ \dots & \dots & \dots & \dots \\ x_{n_k 1}^{(k)} & x_{n_k 2}^{(k)} & \dots & x_{n_k m}^{(k)} \end{matrix} \right\} g_k$$

1. 计算各类均值和离差阵

用 $\bar{x}_j^{(l)}$ 表示第 l 类第 j 个变量的平均数, 则

$$\bar{x}_j^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{ij}^{(l)} \quad j = 1, 2, \dots, m; l = 1, 2, \dots, k \quad (1)$$

用

$$s^{(l)} = \begin{pmatrix} s_{11}^{(l)} & s_{12}^{(l)} & \dots & s_{1m}^{(l)} \\ s_{21}^{(l)} & s_{22}^{(l)} & \dots & s_{2m}^{(l)} \\ \dots & \dots & \dots & \dots \\ s_{m1}^{(l)} & s_{m2}^{(l)} & \dots & s_{mm}^{(l)} \end{pmatrix} \quad l = 1, 2, \dots, k \quad (2)$$

表示第 l 类的离差阵, 它的元素为:

$$s_{ij}^{(l)} = \sum_{a=1}^{n_l} (x_{ai}^{(l)} - \bar{x}_i^{(l)})(x_{aj}^{(l)} - \bar{x}_j^{(l)}) \quad i, j = 1, 2, \dots, m \quad (3)$$

2. 计算总离差阵

用 s 表示各类总的离差阵, 则

$$s = s^{(l)} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \dots & s_{mm} \end{pmatrix}$$

3. 解 k 个方程组

$$\begin{cases} s_{11}c_1^{(l)} + s_{12}c_2^{(l)} + \dots + s_{1m}c_m^{(l)} = \bar{x}_1^{(l)} \\ s_{21}c_1^{(l)} + s_{22}c_2^{(l)} + \dots + s_{2m}c_m^{(l)} = \bar{x}_2^{(l)} \\ \dots \\ s_{m1}c_1^{(l)} + s_{m2}c_2^{(l)} + \dots + s_{mm}c_m^{(l)} = \bar{x}_m^{(l)} \end{cases} \quad l = 1, 2, \dots, k \quad (5)$$

求得 k 组系数

$$\{c_1^{(l)}, c_2^{(l)}, \dots, c_m^{(l)}\}, \quad l = 1, 2, \dots, k \quad (6)$$

它们称为判别系数。

4. 计算常数项 $c_0^{(l)}$

$$c_0^{(l)} = -\frac{1}{2} \sum_{i=1}^m c_i^{(l)} \bar{x}_i^{(l)} \quad l = 1, 2, \dots, k \quad (7)$$

于是得到 k 个判别函数:

$$f^{(l)}(X) = \ln q_l + c_0^{(l)} + c_1^{(l)}x_1 + \dots + c_m^{(l)}x_m \quad (8)$$

其中 q_l 为先验概率, 这里我们可以用样品的频率来代替先验概率。

5. 对原来 $n_1 + n_2 + \dots + n_k$ 个样本进行回报, 看看判别准确与否。判别的规则为: 将任一样本 (它有 m 个指标) $X = (x_1, x_2, \dots, x_m)$ 代入到 (8) 式的 k 个判别函数之中, 看哪一个函数值最大, 比如说 $f^{(2)}(X)$ 最大, 就判该样本属于 g_2 类。如果它本来就属于 g_2 , 那么判断是正确的; 如果它原来不属于 g_2 , 这就产生了误判。当然误判越少越好。

6. 对新的样本进行判断。由于判别函数 $f^{(l)}(X)$ 综合了 k 类物体的各种信息, 因此对一个已知同样指标的新样本可以进行判断, 其判别规则与回报的一样。

当 $k = 2$ 时, 两个判别函数为 $f^{(1)}(X)$ 和 $f^{(2)}(X)$ 可以合成一个, 使用更为方便。本文为了与通常的 Bayes 判别一样, 我们还是使用两个判别函数。

三、材料和结果

1. 材料

我们按雌雄两组分别应用上述的方法,求出各类的判别系数,建立判别函数。对照组为现代人和大猿,雄性共 38 个样本,其中现代人 20 个,大猿 18 个;雌性共 28 个样本,其中现代人 9 个,大猿 19 个。每个样本测了 17 个指标。就指标而言,我们以前五个指标——下颌骨五个项目:犬齿外侧间最大宽(x_1)、下颌联合长(x_2)、下颌在 M_2 中央处高(x_3)、厚(x_4)、 P_2 — M_2 长(x_5)——为一组,又以后十二个指标——即从 I_2 到 M_2 六个牙齿的长和宽(x_6 — x_{17})为另一组。也就是说雌性现代人和大猿前五个指标为一组,后十二个指标为另一组;雄性现代人和大猿前五个指标为一组,后十二个指标为另一组。

判别对象为巨猿,即把在我国柳城巨猿洞发现的三个巨猿下颌骨和在印度发现的一个巨猿下颌骨按各自性别判别其类似于现代人或大猿。对照组和研究组的测量资料见表 1—表 2。

2. 计算结果

对这批样本分别应用 Bayes 判别分析方法求出各类的判别系数。为此先要求出各类指标的总离差阵 $s^{(l)}$ 最后算出各组各类的判别系数 $\{c_1^l, c_2^l, \dots\}$ 整个计算工作是在 DJS-121 计算机上完成的。

由上节公式 (8), 判别函数为:

$$f^{(l)}(x) = \ln q_l + c_0^{(l)} + c_1^{(l)}x_1 + \dots + c_m^{(l)}x_m \quad l = 1, 2, \dots, k$$

其中先验概率 q_l 取成每类出现的样本频率,即用 N 表示 k 类总样本数, n_l 表示第 l 类 g_l 的样本数,取 $q_l = n_l/N$ 。对我们的情况,雌性取 $q_1 = 9/28$, $q_2 = 19/28$;对雄性 $q_1 = 20/38$, $q_2 = 18/38$ 。对某一个样本,将它代入 $f^{(1)}(X)$, $f^{(2)}(X)$, \dots , $f^{(k)}(X)$, 哪个函数值最大,就把它分到那类中。现在我们的问题只有雌雄两类,因此只求 $f^{(1)}(X)$ 和 $f^{(2)}(X)$ 。

对于男人和雄猿的前五个指标(即下颌骨的五个项目)的判别函数为

$$f^{(1)}(X) = \ln(20/38) - 116.4911 + 1.21382x_1 + 0.83952x_2 \\ + 1.60412x_3 + 0.03637x_4 + 4.09084x_5$$

$$f^{(2)}(X) = \ln(18/38) - 330.3586 + 4.03624x_1 + 3.06210x_2 \\ - 0.60359x_3 - 2.22151x_4 + 6.52369x_5$$

把 $X = (x_1, x_2, \dots, x_5)$ 代入上式,求出 $f^{(1)}(X)$ 和 $f^{(2)}(X)$ 的值,然后比较它们的大小。如果 $f^{(1)}(X) > f^{(2)}(X)$ 就类似于人;否则判为类似于猿。对人、猿雄性 38 个样本回报判断全部正确,无一误判。

如将巨猿下颌骨 II 的数据代入得

$$f^{(1)}(X) = 332.695, \quad f^{(2)}(X) = 392.840。$$

显然 $f^{(2)}(X) > f^{(1)}(X)$, 所以应判为类似于猿。

又将巨猿下颌骨 III 的数据代入,得

$$f^{(1)}(X) = 396.33, \quad f^{(2)}(X) = 455.11$$

也是 $f^{(2)}(X) > f^{(1)}(X)$, 所以也应判为猿。

表 1 下颌骨(五项)及其附连牙齿(十二项)的测量(单位: 毫米)

样 本	性 别	犬齿外侧 间最大宽	下颌联合长 (id-gn)	下颌在 M ₂ 中央处		P ₂ -M ₂		I ₁		C ₁		P ₁		P ₂		M ₁		M ₂	
				高	厚	长	长	宽	长	宽	长	宽	长	宽	长	宽	长	宽	长
现代人 (平均数)	♂	32.1600	33.6500	27.6300	17.4000	29.5150	6.4500	6.3800	7.4700	8.0650	7.4450	8.4050	7.4450	8.6750	11.8100	11.0200	11.2950	10.7650	
	♀	32.0667	31.0333	24.9889	17.2000	27.5556	6.4000	6.3000	7.0444	7.6222	7.0778	7.9222	6.9667	8.2889	11.1222	10.6667	10.3444	10.3889	
现生大猿 (平均数)	♂	57.5333	68.0444	37.9444	20.8667	44.3611	8.9500	10.4333	18.4722	14.5000	17.6167	11.9222	11.5444	13.4278	15.4722	13.5833	17.3444	15.3000	
	♀	46.6053	53.4895	32.4947	21.5105	41.8737	8.4947	9.4526	13.1263	10.2105	14.7526	10.4526	10.7474	12.4526	14.7684	12.9579	16.3579	14.4526	
巨猿 I	♀	45.8	73.4	56.7	30.0	52.0	7.8	10.0	11.3	15.1	16.6	14.5	14.6	15.7	17.3	16.1	18.4	16.5	
巨猿 II	♂	53.5	74.0	50.0	33.0	59.0	8.0	10.2	13.4	19.0	17.1	17.3	17.0	16.3	21.2	16.8	21.5	18.7	
巨猿 III	♂	56.0	106.4	81.0	34.0	55.0	7.4	9.3	15.3	19.7	16.4	19.2	16.6	18.0	19.0	18.2	21.2	20.5	
巨猿 IV	?♀	36.0	64.8	53.0	34.0	48.6	5.1	7.2	9.3	16.0	11.8	14.6	13.1	14.5	17.0	14.2	19.1	16.0	
SK-23	?♀	32.8	48.5	35.1	27.8	39.4	7.0	6.6	8.2	8.0	9.1	11.3	10.0	14.2	14.5	14.6	14.9	14.8	
Natron	♂	35.0	50.6	36.9	31.5	46.0	6.0	6.3	7.5	8.0	9.3	13.4	14.0	15.0	15.4	15.3	17.0	16.1	

现代人平均数据古脊椎动物与古人类研究所收藏的华北人和云南人头骨标本测量计算;

现生大猿平均数据 Frayer (1973) 资料计算;

除巨猿 I, II, III 的 M₁ 和 M₂, IV 的 C₁-M₁ 数据引自原始报告外, 巨猿的其他数据测自标本 (I, II, III), 模型 (IV) 或复原部分。在下颌骨一侧的项目, 取右侧数据, 只是在无法获得它时, 才以左侧数据代替之;

粗壮南方古猿资料引自 Frayer (1973)。

表 2 判别结果

样 本	性 别	下颌骨五项指标的 判别函数值		判别结果	牙齿十二项指标的 判别函数值		判别结果
		$f_{II}^{(1)}(x)$	$f_{II}^{(2)}(x)$		$f_{II}^{(1)}(x)$	$f_{II}^{(2)}(x)$	
巨猿 II	♂	$f_{II}^{(1)}(x)$ 332.69	$f_{II}^{(2)}(x)$ 392.84	大猿	$f_{II}^{(1)}(x)$ 398.19	$f_{II}^{(2)}(x)$ 332.15	人
巨猿 III	♂	$f_{III}^{(1)}(x)$ 396.33	$f_{III}^{(2)}(x)$ 455.11	大猿	$f_{III}^{(1)}(x)$ 512.94	$f_{III}^{(2)}(x)$ 409.95	人
Natron	♂	$f^{(1)}(x)$ 216.02	$f^{(2)}(x)$ 173.06	人	$f^{(1)}(x)$ 397.95	$f^{(2)}(x)$ 263.67	人
巨猿 I	♀	$f_I^{(1)}(x)$ 481.54	$f_I^{(2)}(x)$ 575.84	大猿	$f_I^{(1)}(x)$ 581.96	$f_I^{(2)}(x)$ 576.11	人
巨猿 IV	?♀	$f_{IV}^{(1)}(x)$ 385.87	$f_{IV}^{(2)}(x)$ 402.65	大猿	$f_{IV}^{(1)}(x)$ 589.38	$f_{IV}^{(2)}(x)$ 501.02	人
SK-23	?♀	$f^{(1)}(x)$ 287.89	$f^{(2)}(x)$ 278.99	人	$f^{(1)}(x)$ 351.98	$f^{(2)}(x)$ 249.34	人

于是两个雄性巨猿下颌骨均判为类似大猿。

类似的,用牙齿长度和宽度的 12 个指标建立判别函数后,用它们来判别巨猿 II、III,说明它们的牙齿类似于人。

同样,我们分别建立了雌性组的下颌骨和牙齿的判别函数,也作了回报,同样对巨猿 I 和 IV(即印度比拉斯普巨猿)作了判别,其结果与雄性组完全一样。即下颌骨类似于大猿;牙齿类似于人。

最后,我们用粗壮南方古猿下颌骨和牙齿相同的指标作对照,以验检巨猿的判别结果。雄性的 Natron 下颌骨、据原报告称可能是雌性的 Swartkrans 的 SK-23 下颌骨的每项判别全为类似于人。全部结果列在表 2。

根据以上进行的判别分析,我们得出了一致的结果:就 I_2-M_2 六个牙齿的长度、宽度共 12 项指标总体而言,巨猿类似于现代人(而不是大猿)。剩下的问题是,产生这种类似的原因到底是亲缘关系,还是趋同进化?王令红、吴茂霖、张银运(1978,第 71—72 页)已经谈到,现有证据对于说明巨猿与人类的亲缘关系更为有利。这里不再赘述。

因此, Bayes 判别结果,对于解释巨猿在从猿到人的进化过程中的地位问题颇为有益,具体地说,虽然现已发现的两种巨猿是绝灭旁支,但它们是后期发展的结果,早期祖先应该是从猿到人的过渡类型,在它们脱离这个进化主支之前,已经经历了从树居到下地生活的变化。食物来源的变化使这种过渡类型生物的变化最早、最直接地表现在牙齿形态及其大小上。

(1980 年 6 月 26 日收稿)

参 考 文 献

- 王令红、吴茂霖、张银运, 1978: 论巨猿与人类的关系, 古人类论文集——纪念恩格斯«劳动在从猿到人转变过程中的作用»写作一百周年报告会论文汇编, 67—76, 科学出版社。
- 吴汝康, 1962: 巨猿下颌骨和牙齿化石, 中国古生物志新丁种第十一号, 1—94。
- Anderson, T. W., 1958: An Introduction to Multivariate Statistical Analysis, John Wiley.
- Corruccini, R. S., 1975: Multivariate Analysis of *Gigantopithecus* Mandibles, *Am. J. Phys. Anthropol.*, **42**: 167—170.
- Frayer, D. W., 1973: Gigantopithecus and its Relationship to *Australopithecus*, *Am. J. Phys. Anthropol.*, **39**: 413—426.
- Robinson, J. T. and Karen Steudel, 1973: Multivariate Discriminant Analysis of Dental Date Bearing on Early Hominid Affinities. *J. Hum. Evol.*, **2**: 509—527.
- Simons, E. L. and S. R. K. Chopra 1969: *Gigantopithecus* (Pongidae, Hominoidea) a New Species from North India, *Postilla Yale Peabody Museum*, No. 138: 1—18.

AN APPLICATION OF BAYES DISCRIMINANT ANALYSIS IN DETERMINING THE SYSTEMATIC POSITION OF *GIGANTOPITHECUS*

Wang Linghong

(Institute of Vertebrate Paleontology and Paleoanthropology, Academia Sinica)

Fang Kaitai Zheng Yuying

(Institute of Applied Mathematics, Academia Sinica)

Abstract

The systematic position of *Gigantopithecus* in Hominoidea has been a subject of heated discussion among anthropologists for a long time. Several foreign scholars have discussed this problem from the view-point of numerical taxonomy. We try to reexamine the status of *Gigantopithecus* by means of multivariate discriminant analysis. In this paper we introduce the method of Bayes discriminant analysis and show the result we have obtained through its application.

Gigantopithecus was compared to both modern man and the gorilla to determine whether it resembles the former or the latter. Our discriminant analysis is composed of two parts: (1) mandibles (5 variables: bicanine breadth, symphyseal height, corpus height and breadth at M_2 , P_2 — M_2 chord); (2) teeth (12 variables: length and breadth of I_2 through M_2). Male and female specimens are dealt with respectively.

The result of discriminant analysis shows that the mandibles of both male and female *Gigantopithecus* are similar to those of gorilla, but the teeth, in contrast, more closely resemble those of modern man.

The resemblance between *Gigantopithecus* and modern man may be regarded as an affinity, rather than a convergence. So *Gigantopithecus* should have ancestral status as a creature translated from ape to man, which was no longer arboreal, but lived on the ground. The change of food source led to the change of the dental morphology and its size firstly and directly.