利用 SAS 软件对 271 株肠杆菌进行数值分类初探

郭秀花 李爱国 徐桂永 于化路 徐鹏辉

(解放军北京医学高等专科学校 北京 100071)

摘要 利用 SAS 软件对肠杆菌科中的 90 株标准菌进行系统聚类，去掉独立菌再结合判别分析的应用条件，建立了肠杆菌关于 7 类属(或种)的筛选变量与不筛选变量的两种判别函数。对 181 株临床菌用两种判别函数进行判别，判得结果的一致率为 87.57%。对标准菌进行回代，与原分类情况作比较，结果是：不筛选变量时符合率为 89.41%，筛选变量时符合率为 97.65%。

关键词 肠杆菌，SAS 软件，聚类分析

分类号 Q939 **文献识别码** B **文章编号** ISSN-0253-2654(1999)-01-50-52

MAKING NUMERICAL CLASSIFICATION AND IDENTIFICATION TO 271 *ENTEROBACTERIACEAES* BY SAS SOFTWARE

Guo Xiuhua, Li Aiguo, Xu Guiyong, Yu Hualu, Xu Penghui

(*Beijing Medical College of PLA, Beijing* 100071)

Abstract SAS software is the one of the best software In the world. This article makes hierarchical cluster to 90 standard *enterobacteriaceaes* by the software. We have obtained discriminant functions of screening variable and no screening variable when without simple *enterobacteriaceae* and in the light of the condition of discriminant analysis. Consistent rate is 87.57% when discriminanting 181 clinical *enterobacteriaceaes*; coincident rates of no screening variable and no screening variable are 97.65% and 89.41% respectively.

Key words *Enterobacteriaceae*, SAS software, Cluster analysis

微生物中细菌的分类方法有传统分类法、数值分类法、分子遗传学分类法及化学分类法等[1]。随着医学的迅猛发展，利用计算机对细菌进行数值分类，已越来越受到人们的重视。数值分类的方法很多[2]，本文利用 SAS 软件对 271 株肠杆菌进行数值分类作了初步探讨，现报道如下。

1 材料

1.1 指标

为建立快速稳定的肠杆菌细菌的鉴定方法，我们在厌氧菌脱氢酶研究的基础上并参考有关文献，选取了 20 项脱氢酶快速反应指标[3~5]：5 种氨基酸、5 种有机酸、2 种有机醇、6 种双糖和 2 种单糖，按 0.03mL 浓度组装 40 孔酶联反应板。

1.2 菌种

由天坛生物检定所菌种保藏中心及军事医学科学院提供肠杆菌标准菌株 90 株，具体见表 1。

解放军 302 医院、304 医院、海军总医院检验科细菌室提供了许多临床菌株，本文采用了 181 株，包含的属及种的情况是：除含有上述的 9 个属、18 种外，还有 39 株其它菌。将 271 株细菌接种半固体并置 −74℃冰箱保存，使用时，首先在普通琼脂上传 2 代复苏，再转种在特制脱氢酶诱导培养平板上培养 18～24h，取菌落制

1997-12-04收稿，1998-04-20修回

表1　90株标准菌所代表的属及种的情况

	属名								
	大肠埃希菌	沙门菌	志贺菌	枸橼酸菌	变形杆菌	沙雷菌	摩根菌	肠杆菌	雷极菌
种数	1	6	4	1	2	1	1	1	1
菌数	12	19	22	3	21	5	5	1	2

备 0.025M 的 PBS 菌悬液 4.0mL、浊度为 4 个 McFarland 标准(约 12 亿 / mL)。

1.3　酶学反应

参照文献[6]每孔加无菌液体石蜡一滴以便阻止空气氧的干扰作用，37℃反应 30min，红色为脱氢酶阳性，本底色为阴性反应[7-8]。

1.4　收集数据

在选取的 20 项指标中，每 1 株细菌在每项指标下的反应是阳性的，输入微机时记为 1，阴性的记为 0。每株菌根据稳定情况重复做 2 至 15 次实验，取其均值(即出现阳性的次数之和，除以实验次数所得的频率)作为在相应指标下的原始数据，271 株肠杆菌在 20 项指标下的资料见表 2。

表2　271株肠杆菌20项指标下的资料

	X±S		
	标准菌	临床菌	合　计
例　数	90	181	271
色氨酸X_1	0.2472±0.4338	0.0946±0.2911	0.1442±0.3506
香草醛X_2	0.0449±0.2084	0.1230±0.3236	0.0976±0.2931
肌　醇X_3	0.1236±0.3310	0.1108±0.3071	0.1150±0.3145
亮氨酸X_4	0.2562±0.4369	0.2608±0.4344	0.2593±0.4345
苏氨酸X_5	0.2146±0.3913	0.2514±0.4271	0.2394±0.4154
赖氨酸X_6	0.3558±0.4495	0.2684±0.4260	0.2968±0.4359
精氨酸X_7	0.0154±0.0958	0.0878±0.2722	0.0643±0.2325
乙酸钠X_8	0.5349±0.4606	0.6838±0.4559	0.6354±0.4619
阿拉伯糖X_9	0.5004±0.4727	0.6900±0.4572	0.6282±0.4699
丙二酸钠X_{10}	0.0073±0.0396	0.0216±0.1362	0.0170±0.1143
葡萄糖X_{11}	0.9865±0.1079	1.0000±0.0000	0.9956±0.0616
乳　糖X_{12}	0.1685±0.3765	0.4838±0.4997	0.3814±0.4857
麦牙糖X_{13}	0.9865±0.1079	0.9811±0.1315	0.9828±0.1242
甘露醇X_{14}	0.6270±0.4845	0.8608±0.3397	0.7849±0.4068
蔗　糖X_{15}	0.1528±0.3555	0.4122±0.4884	0.3279±0.4651
纤维二糖X_{16}	0.0901±0.2648	0.4523±0.4924	0.3346±0.4635
海藻糖X_{17}	0.7989±0.3950	0.8658±0.3367	0.8441±0.3573
甜　醇X_{18}	0.0857±0.2399	0.0068±0.0757	0.0324±0.1542
谷氨酸X_{19}	0.7520±0.4129	0.7068±0.4456	0.7215±0.4351
葡萄糖酸X_{20}	0.5401±0.4445	0.6496±0.4700	0.6140±0.4639

2　方法及结果

2.1　标准菌株的聚类

SAS 软件聚类分析提供了五个过程：Cluster、Fastclus、Varclus、Tree 及 Aceclus 过程[9]。我们采用了 Cluster 系统聚类过程对 90 株标准菌株进行了聚类。

Cluster 过程中有 11 种聚类形式(即距离的定义方法不同)：AVERAGE(类平均法)、CENTROID(重心分量法)、COMPLETE(最长距离法)、DENSITY(非参数概率密度估计法)、EML(最大似然法)、FLEXIBLE(flexible-beta 法)、MCQUITTY(Mcquitty 的相似分析法)、MEDIAN(中位数法)、SINGLE(最短距离法)、TWOSTAGE(两阶段密度法)和 WARD(Ward 最小方差法)。针对 11 种聚类形式的特点，对 90 株标准菌经类平均法、重心分量法、最小方差法、最短距离法、中位数法等几种形式进行聚类。结合专业上标准菌来自的属及种的情况及统计上聚类结果较佳标准，得知用类平均法分成 13 类比较理想。

在分得的 13 类中有 5 类是独立菌，菌株的编号分别为：9、57、78、80、85，另外 8 类的分类结果见表 3。

表3　进行Cluster聚类后非独立菌的分类结果

类别	例数	类内菌株编号
1	30	1—5 15 22 24 32 36—47 50 62—67 79 86
2	26	14 21 23 28—31 33 34 35 58—61 68—77 89 90
3	5	10 11 12 20 48
4	10	6 7 25 49 54 56 82 83 87 88
5	2	13 19
6	7	8 26 27 51 53 81 84
7	3	16 17 18
8	2	52 55

为了用判别分析方法对临床细菌进行鉴定，按照判别分析的要求，每类例数不能少于 3，将距离系数最靠近的 4 和 5 两类合并、57 号独立菌加入到第 8 类，这样进行微调后就得到了 7 类，各类在 20 个指标下的均数(即频率矩阵)见表 4。

表4 7类在20个指标下的均数

指标	类别编号						
名称	1	2	3	4	5	6	7
X_1	0.0000	0.0000	0.0000	0.9167	1.0000	0.0000	1.0000
X_2	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
X_3	0.0000	0.0000	0.2000	0.0000	1.0000	0.0000	0.3333
X_4	0.0000	0.0000	0.0000	1.0000	1.0000	0.0000	1.0000
X_5	0.0533	0.0000	0.3200	0.7750	0.6000	0.0000	0.8000
X_6	0.3067	0.0160	0.8000	0.8058	0.7143	0.0000	0.8000
X_7	0.0000	0.0000	0.0000	0.0142	0.0000	0.0000	0.1333
X_8	0.8557	0.0160	1.0000	0.3367	0.7429	0.8333	0.3333
X_9	0.9240	0.1600	1.0000	0.2392	0.0571	0.3333	0.5167
X_{10}	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2167
X_{11}	1.0000	0.9920	1.0000	1.0000	1.0000	1.0000	1.0000
X_{12}	0.3333	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
X_{13}	1.0000	0.9920	1.0000	1.0000	1.0000	1.0000	1.0000
X_{14}	1.0000	0.5920	1.0000	0.0833	0.0000	1.0000	0.0000
X_{15}	0.0000	0.0400	0.4000	0.0167	1.0000	0.3333	0.4667
X_{16}	0.0340	0.0000	0.9200	0.1000	0.0286	0.0000	0.0000
X_{17}	1.0000	0.3920	1.0000	0.9000	1.0000	1.0000	0.8333
X_{18}	0.0367	0.0000	0.0800	0.0000	0.4757	0.0000	0.2667
X_{19}	0.8443	0.3520	0.9600	1.0000	1.0000	1.0000	1.0000
X_{20}	0.7640	0.1080	0.9600	0.4475	0.5900	1.0000	0.7833
属或种	埃希与沙门	志 贺*	枸橼酸*	奇异变形*	普通变形	粘质沙雷	摩根菌

* 指本类的属(或种)中有其它属(或种)的菌株

2.2 对临床菌进行归类判别

判别分析是根据已掌握的一批分类明确的样品，建立较好的判别函数，使产生错判的例数最少，进而对给定的各个新样品，判断它的归类。对于定量资料的判别分析，SAS 软件提供了考虑变量筛选的逐步判别分析 STEPDISC 过程及不考虑变量筛选的一般判别分析 DISCRIM 过程。

对于 86 株标准菌按照表 2 的分类情况，分别进行筛选变量及不筛选变量的判别分析，建立判别函数。不筛选变量的一般判别分析，各个自变量对 7 类总体鉴别能力的多元检验结果 $P < 0.0001$，说明所有变量建立的判别函数具有极显著的判别能力，对标准菌进行回代判别，符合率[9](指原各菌株所在的类别与经判别分析后所在的类别相符合的细菌株数除以被判断的细菌总数)为 89.41%。而考虑变量筛选的逐步判别分析，各个自变量对 7 类总体鉴别能力的多元检验结果 $P < 0.0001$，说明筛选变量后的较少变量建立的判别函数具有极显著的判别能力，对标准菌进行回代判别，符合率为 97.65%。因此这两种判别函数均可用来鉴别上述 7 类肠杆菌中的各属(或种)。

用两种判别函数对 181 株临床菌进行归类判别，判断结果的一致率为 87.57%。筛选变量时 1 至 7 类中各属(或种)的例数依次为 38、36、44、28、11、23、1。

3 讨论

目前对细菌的数值分类，国内大多采用的是专用软件包，能处理的最大数据例数是 255[10]，而国际流行的统计软件包能处理的数据例数没有限制。如何对肠杆菌科细菌利用 SAS 软件进行数值分类，我们从分类方法角度，进行了初步的探讨。

本研究的全部资料输入 586 微机，建 FoxBase 库后，调入 SAS 软件的 Cluster 过程并自编树枝图对标准菌进行聚类(由于篇幅所限树枝图略)，将专业知识及聚类最佳判断标准有机地结合起来，才能得到比较理

(下转第 67 页)

(上接第 52 页)

想的数值分类结果。

根据聚类情况建立判别函数后对标准菌进行回代，并非指标越多越好，当指标间存在共线性关系互相影响时，其鉴别能力反而降低(筛选变量与不筛选变量时的符合率分别为97.65%和89.41%)。另外，尽量减少指标也可以减少收集数据的工作量[11]。

利用聚类分析对标准菌株进行聚类，建立判别函数后再对未知菌进行归类判别，需要有足够多的标准菌资料，否则会出现有的未知菌本不属于用标准菌所分的类别，而被错判的情况。如39株临床菌无对应的标准菌株，而用判别分析均被判为7类中的某一类了。

我们将进一步收集肠杆菌科细菌的标准菌株，完善SAS软件进行肠杆菌科细菌的数值分类方法，建立较全面的肠杆菌科细菌的数值分类鉴定模型。

致谢 海军总医院胡九茹、杜连荣，304医院常东及302医院黄上碗同志提供了大量的菌株，在此表示感谢。

(下转第 76 页)

(上接第 67 页)

参 考 文 献

[1] 林万明. 细菌分子遗传分类鉴定法. 上海:上海科学技术出版社,1990.

[2] 郭秀花,李爱国,冯 丹等. 微生物学通报,1996,23(3):188～190

[3] 李爱国,何道生. 中国微生态杂志,1991,3(4):93～96.

[4] Holmes B. J Clin Microbiol. 1989, 27(5):1027～1030.

[5] Gharbia SE, Haroun NS. J Generl Microbiol, 1988; 134:327～329.

[6] 李爱国,何道生,林万明等. 中国微生态杂志,1992,3(4):9～13.

[7] 李爱国,郭秀花,刘淑珍等. 中国人民解放军北京医高专学报,1996,5(2):6～9.

[8] 胡良平主编. 现代统计学与 SAS 应用. 北京:军事医学科学出版社,1996,336～366.

[9] 许晓东,陈文新. 微生物学通报,1996,23(3):131～134.

[10] 郭秀花,李爱国,徐桂永等. 微生物学通报,1998,25(1):24～26.