20 项肠杆菌酶学指标的筛选

郭秀花 李爱国 徐桂永 张德舜 朴相浩

(中国人民解放军北京医学高等专科学校 北京 100071)

摘要 肠杆菌科酶法快速分类系统的研制工作中，初选了 20 项酶学指标，利用 SAS 软件的主成分分析和变量聚类分析法，结合变异系数法及相关系数法，按照专业知识的特点，对 271 株肠杆菌的实验数据结果，筛选出了主要的 10 项指标。此研究为提取主要指标，并对指标加权作进一步处理提供了科学依据。

关键词 肠杆菌，指标，SAS 软件

分类号 Q939

选择指标是医学研究中经常会遇到的问题。指标并非越多越好，指标要有客观性、敏感性和特异性。指标过多时，不但收集数据的工作量大，而且给分析问题带来不便，有时会使问题复杂化。建立有效的指标体系，对反映事物的本质特征十分重要。细菌数值分类的各种方法[1]，也要在多种合理的指标体系下进行。对 271 株肠杆菌进行实验，运用多元统计分析中的主成分分析，聚类分析，再结合变异系数法，相关系数法[2]，从较多的原始指标中筛选了 10 项指标。其结果基本反映了原来指标中所包含的主要的共同信息。为进一步加权数值分类、提高符合率、减少工作量和提高经济效益，做了有意的探索。

1 材料与方法

1.1 材料

在"肠杆菌科细菌酶法快速分类系统"的课题研究中，建立了细菌细胞结构的脱氢酶检测方法，并以完整细胞脱氢酶活性和部分水解酶活性的脱氢酶反应作为细菌分类标志。初选了 20 项指标，共做了 271 株(标准菌 90 株，临床菌 181 株)肠杆菌科细菌的实验结果。每一株细菌在每项指标下的反应是阳性输入微机时记为 1、阴性记为 0，每菌株根据稳定情况重复做 2～15 次实验，取其均值作为在相应指标下的原始数据，资料如表 1 所示。

1.2 筛选方法

1.2.1 主成分分析[3,4]：采用多个定量(数值)变量间相关性的一种多元统计分析方法。它是研究如何通过少数几个主分量(即原始变量的线性组合)来解释多变量的方差 – 协方差结构。即是根据累积贡献率的大小导出少数几个主分量，它们尽可能地保留了原始变量的信息，且彼此间又不相关。然后进行最大方差旋转，挑选因子负荷绝对值大的指标。

1.2.2 变量聚类[5,6]：SAS 软件聚类分析，提供了五个过程：Cluster、Fastclus、Varclus、Tree、Aceclus 过程。我们采用了 Varclus 过程对 271 株肠杆菌进行了聚类。Varclus 过程是基于相关矩阵或协方差阵对变量进行分割聚类或系统聚类，对于每一类，过程都要计算其第一主成分或者重心成分，并且过程尽可能使各个类中类成分所含的信息之和达到最大。

1.2.3 变异系数法：是从指标的敏感性角度挑选指标，变异系数 CV(见表 1)太小，在肠杆菌分类时分辨率就低。所以，从指标的各类中尽量剔除变异系数及实际意义较小者。

1.2.4 相关系数法：是从指标的代表性与独立性角度挑选指标。主成分分析的中间过程提供

1997-01-10收稿

© 中国科学院微生物研究所期刊联合编辑部 http://journals.im.ac.cn

表1　271株肠杆菌酶法20项指标下的资料

指标	均数x	标准差S	变异系数CV(%)	指标	均数x	标准差S	变异系数CV(%)
色氨酸X_1	0.1442	0.3506	243.1	葡萄糖X_{11}	0.9956	0.0616	6.2
香草醛X_2	0.0976	0.2931	300.3	乳　糖X_{12}	0.3814	0.4857	127.3
肌　醇X_3	0.1150	0.3145	273.5	麦芽糖X_{13}	0.9828	0.1242	12.6
亮氨酸X_4	0.2593	0.4345	167.6	甘露醇X_{14}	0.7849	0.4068	51.8
苏氨酸X_5	0.2394	0.4154	173.4	蔗　糖X_{15}	0.3279	0.4651	141.8
赖氨酸X_6	0.2968	0.4349	146.5	纤维二糖X_{16}	0.3346	0.4635	138.5
精氨酸X_7	0.0643	0.2325	361.6	海藻糖X_{17}	0.8441	0.3573	42.3
乙酸钠X_8	0.6354	0.4619	72.7	甜　醇X_{18}	0.0324	0.1542	475.9
阿拉伯糖X_9	0.6282	0.4699	74.8	谷氨酸X_{19}	0.7215	0.4351	60.3
丙二酸钠X_{10}	0.0170	0.1143	672.3	葡萄糖酸X_{20}	0.6140	0.4639	75.6

了 20 项指标的相关矩阵，对变量聚类中是一类的变量，尽量剔除相关系数较大而实际意义较小者。

2　结　果

2.1　主成分分析

利用 SAS 中的 proc princomp 过程进行主成分分析，271 株肠杆菌资料主成分分析的相关矩阵的特征值如表 2 所示。

表2　相关矩阵的特征值

主成分	方差	方差之差	贡献率 (%)	累积贡献率 (%)
Z1	4.9681	2.0329	0.2484	0.2484
Z2	2.9353	1.2161	0.1468	0.3952
Z3	1.7192	0.3581	0.0860	0.4811
Z4	1.3611	0.2205	0.0681	0.5492
Z5	1.1406	0.0845	0.0570	0.6062
Z6	1.0561	0.0406	0.0528	0.6590
Z7	1.0155	0.1390	0.0508	0.7098
Z8	0.8765	0.0744	0.0438	0.7536
…	…	…	…	…
Z20	0.2044		0.0102	1.0000

结合常规细菌的分类特点，选取了大于 75% 的累积贡献率，包含有八个主成分，各主成分的组合系数就是相关矩阵的特征向量。从特征向量表(略)中得知：第一主成分负荷较大的指标是：X_8、X_{12}、X_{15}、X_{16}；第二主成分负荷较大的指标是：X_1、X_{14}、X_4；第三主成分负荷较大的指标是：X_{11}、X_{13}；第四至第八个主成分负荷较大的指标依次是：X_3、X_{18}、X_7、X_{10}、X_2。

2.2　Varclus 聚类

本文调用了 4 个 Varclus 聚类过程，具体是：proc varclus、proc varclus centroid、proc varclus hi maxc = 15 和 proc varclus centroid maxc = 15。结合专业及统计学知识，综合判断比较，第二个过程中聚成 10 类较好。此时，把全部 20 个指标聚成 10 类，能解释的方差为 15.4803，占总方差的 77.40%，分类情况如表 3 所示。

表3　Varclus centroid过程聚10类情况

序号	例数	类中变量的类方差	由类成分所解释的方差	解释的方差所占比例	类中指标名称
1	4	4.0000	2.7955	0.6989	X_8 X_9 X_{19} X_{20}
2	3	3.0000	2.1085	0.7028	X_1 X_4 X_5
3	2	2.0000	1.4885	0.7442	X_{11} X_{13}
4	1	1.0000	1.0000	1.0000	X_{10}
5	2	2.0000	1.4081	0.7041	X_6 X_7
6	1	1.0000	1.0000	1.0000	X_{17}
7	2	2.0000	1.4047	0.7024	X_3 X_{18}
8	1	1.0000	1.0000	1.0000	X_2
9	3	3.0000	2.2651	0.7550	X_{12} X_{15} X_{16}
10	1	1.0000	1.0000	1.0000	X_{14}

2.3　变异系数法及相关系数法

由表 1 可知，变异系数 CV 较小者是：X_{12}、X_{13}、X_{17}、X_{14}、X_{19}、X_8、X_9、X_{20}；由主成分分析的中间过程中，得知相关系数较大的变量依次是：X_8

© 中国科学院微生物研究所期刊联合编辑部　http://journals.im.ac.cn

和 X_{20}、X_{12}和 X_{16}、X_1和 X_4、X_{15}和 X_{16}、X_9和 X_{20}、X_8和 X_9、X_{19}和 X_{20}、X_9和 X_{12}、X_1和 X_{15}、X_8和 X_{19}，…。从聚成 10 类的结果中，对每一类挑选主要指标，其方法是：结合主成分分析结果的主要指标，利用变异系数法及相关系数法筛选指标总的原则，再从专业知识方面综合比较分析，最后筛选出了 10 个主要指标，即：X_1、X_2、X_3、X_7、X_8、X_{10}、X_{12}、X_{13}、X_{14}、X_{17}。

3 讨 论

SAS 软件是美国 SAS 公司于 1966 年开始研制的，版本不断更新，它是当前国际上最流行、最有权威性的统计分析软件之一。本文采用的是 PC / SAS 6.04 版，研究结果与常规细菌分类结果基本相符，比较理想，为利用 SAS 软件迅速、准确地进行指标筛选作了有益的尝试。

综合运用主成分分析、变量聚类、变异系数法和相关系数法，对 271 株肠杆菌 20 项酶学指标下的资料，进行指标的不同侧面筛选，入选的指标占原指标的 50%，这 10 项指标分别是：色氨酸、香草醛、肌醇、精氨酸、乙酸钠、丙二酸钠、乳糖、麦芽糖、甘露醇、海藻糖。为在肠杆菌科脱氢酶法快速分类系统中，选取主要指标，加权分类鉴定，减少工作量，提高经济效益，提供了科学的依据。

统计分析的结果要与专业实际相结合。比如 SAS 软件提供了不同的聚类过程，每个过程又可以聚成数目不同的类，到底应选取何过程下的几类聚类结果，一方面要从统计上考虑是最优的情况，另一方面又要考虑肠杆菌酶学的 20 项指标应分几类，哪些指标重要。只有两者结合起来，才能获得满意的结果。

参 考 文 献

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

[2] 殷银康，田凤调. 中国卫生事业管理，1994，1：48～51.

[3] 高惠璇主编. SAS 系统使用手册(四). 北京：北京大学出版社，1995，112～162.

[4] 张菊英，倪宗璧. 现代预防医学，1994，21(1)：5～8.

[5] Everitt B S. Cluster Analysis. Second Edition. London: Heineman Educational Books Ltd, 1980.

[6] 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，1996，349～354.

SCREENING PRINCIPAL INDEXES FROM 20 ENZYMATIC INDEXES FOR THE MEMBERS OF THE FAMILY ENTEROBACTERIACEAE

Guo Xiuhua Li Aiguo Xu Guiyong Zhang Deshun Pu Xianghao

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

Abstract In the esarching a rapid classification system of the enzymes for the members of the family enterobacteriaceae, We have selected 20 enzymatic indexes. Furthermore, According to professional knowledge we have screened 10 indexes for experiment results of 271 enterobacteriaceaes of by using principal component analysis and variable cluster analysis of SAS software, incidentally by using coefficient of variation method and correlation coefficient method The Study has provided scientific basis for further reseach into principal indexes and giving weight to the indexes

Key words Enterobacteriaceae, Index, SAS software

© 中国科学院微生物研究所期刊联合编辑部 http://journals.im.ac.cn